Capstone Project – Applied Statistics Interview Grind

Problem Statement: Write the Solutions to any 50 Interview Questions out of 80 total questions.

1. What is a vector in mathematics?

2. How is a vector different from a scalar?

3. What are the different operations that can be performed on vectors?

4. How can vectors be multiplied by a scalar?

5. What is the magnitude of a vector?

6. How can the direction of a vector be determined?

7. What is the difference between a square matrix and a rectangular matrix?

8. What is a basis in linear algebra?

9. What is a linear transformation in linear algebra?

10. What is an eigenvector in linear algebra?

11. What is the gradient in machine learning?

12. What is Backpropogation in machine learning?

13. What is the concept of a derivative in calculus?

14. How are partial derivatives used in machine learning?

15. What is probability theory?

16. What are the primary components of probability theory?

17. What is conditional probability, and how is it calculated?

18. What is Bayes theorem, and how is it used?

19. What is a random variable, and how is it different from a regular variable?

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

21. What is the central limit theorem, and how is it used?

22. What is the difference between discrete and continuous probability distributions?

23. What are some common measures of central tendency, and how are they calculated?

24. What is the purpose of using percentiles and quartiles in data summarization?

25. How do you detect and treat outliers in a dataset?

26. How do you use the central limit theorem to approximate a discrete probability distribution?

27. How do you test the goodness of fit of a discrete probability distribution?

28. What is a joint probability distribution?

29. How do you calculate the joint probability distribution?

30. What is the difference between a joint probability distribution and a marginal probability distribution?

31. What is the covariance of a joint probability distribution?

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

33. What is the relationship between the correlation coefficient and the covariance of a joint probability distribution?

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

35. What are the different sampling methods commonly used in statistical inference?

36. What is the central limit theorem, and why is it important in statistical inference?

37. What is the difference between parameter estimation and hypothesis testing?

38. What is the p-value in hypothesis testing?

39. What is confidence interval estimation?

40. What are Type I and Type II errors in hypothesis testing?

41. What is the difference between correlation and causation?

42. How is a confidence interval defined in statistics?

43. What does the confidence level represent in a confidence interval?

44. What is hypothesis testing in statistics?

45. What is the purpose of a null hypothesis in hypothesis testing?

46. What is the difference between a one-tailed and a two-tailed test?

47. What is experiment design, and why is it important?

48. What are the key elements to consider when designing an experiment?

49. How can sample size determination affect experiment design?

50. What are some strategies to mitigate potential sources of bias in experiment design?

51. What is the geometric interpretation of the dot product?

52. What is the geometric interpretation of the cross-product?

53. How are optimization algorithms with calculus used in training deep learning models?

54. What are observational and experimental data in statistics?

55. How are confidence tests and hypothesis tests similar? How are they different?

56. What is the left-skewed distribution and the right-skewed distribution?

57. What is Bessel's correction?

58. What is kurtosis?

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

60. What is the difference between Descriptive and Inferential Statistics?

61. 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?

62. 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?

63. What is the meaning of degrees of freedom (DF) in statistics?

64. 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)?

65. What is the empirical rule in Statistics?

66. What is the relationship between sample size and power in hypothesis testing?

67. Can you perform hypothesis testing with non-parametric methods?

68. What factors affect the width of a confidence interval?

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

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

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

72. What is the relationship between the margin of error and confidence interval?

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

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

75. What is a Chi-Square test?

76. What is a t-test?

77. What is the ANOVA test?

78. How is hypothesis testing utilised in A/B testing for marketing campaigns?

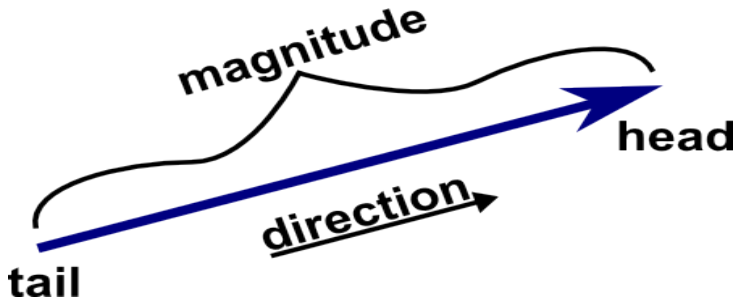
79. What is the difference between one-tailed and two tailed t-tests?

80. What is an inlier?

Solutions to 50 Interview Questions

Q1.What is a vector in mathematics?

Definition- A vector is an object that has both a magnitude and a direction. Geometrically, we can picture a vector as a directed line segment, whose length is the magnitude of the vector and with an arrow indicating the direction. The direction of the vector is from its tail to its head. We denote vectors using boldface as in **a** or **b**. While writing vectors using arrows as in $\vec{a}$ or $\vec{b}$. We denote the magnitude of the vector a by ‖a‖.



Types of Vectors

<u>There are 10 types of vectors</u>:

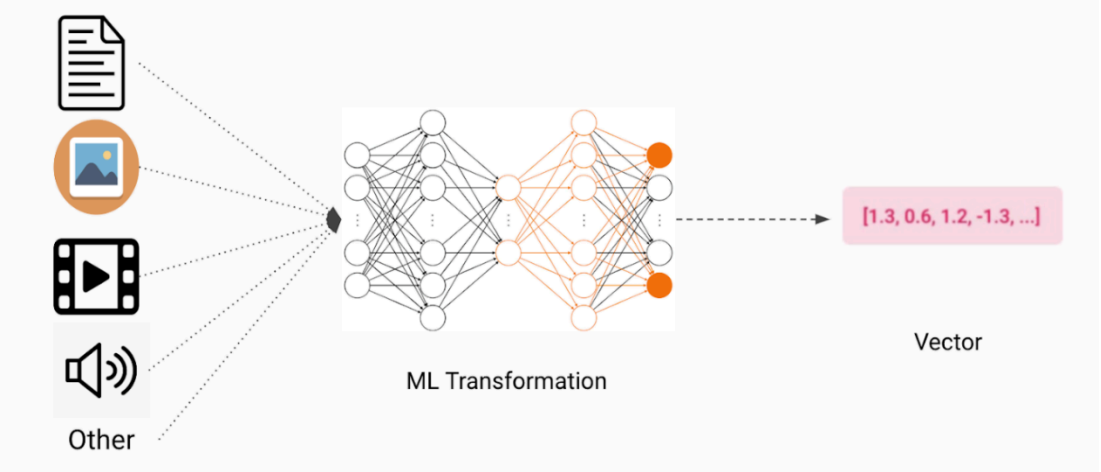
- When the magnitude is 1, it is called a ***unit vector***.
- If two vectors start from the same point, it is called ***Co-initial vectors***.
- If two vectors lie in the same plane, it is called ***Co-planar vectors***.
- If two vectors have equal magnitude and travel in the same direction, it is called ***equal vectors.***
- If two vectors have same magnitude but opposite direction, it is called ***negative vectors.***
- If the starting and ending position of a vector is the same, it is called ***zero vector***.
- If a vector denotes its position in a three dimensional system, it is called ***position vector.***
- If two vectors have same direction it is called like vectors and if it is in opposite direction, it is called ***unlike vectors***.
- If two vectors lie in same line or are parallel to each other, it is called ***collinear vectors.***
- If the position of a vector is displaced from one point to another, it is called ***displacement vector.***

Practical Application of Vector in Everyday Life

Vectors are mathematical objects that contain both magnitude and direction, and they can be represented by the directed line segments (lines having directions) whose lengths are their magnitude. It is used to describe the movement of an object from one point to another.

In Data Science, vectors are used to represent numeric characteristics, called features, of an object in a mathematical and easily analysable way. Vectors are essential for many different areas of machine learning and pattern processing. A vector is a data structure with at least two components: magnitude and direction. It is most commonly used in machine learning to represent the data in the most optimized and organized way.

In machine learning, while training any model, if our dataset contains the images and text, they are first converted into numbers and then stored in the form of vectors and matrices to represent these data, i.e., the very first step in building a machine learning model is vectorizing the data. Once the data is vectorized, you can easily use different linear algebra operations (or tools) to perform model training, data augmentation.

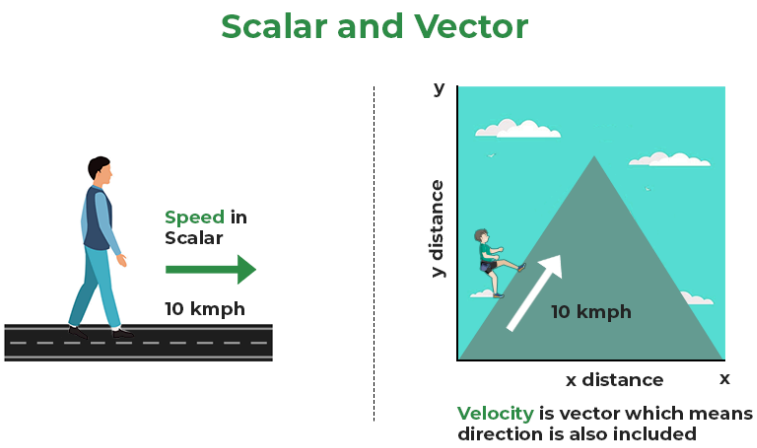


Conclusion

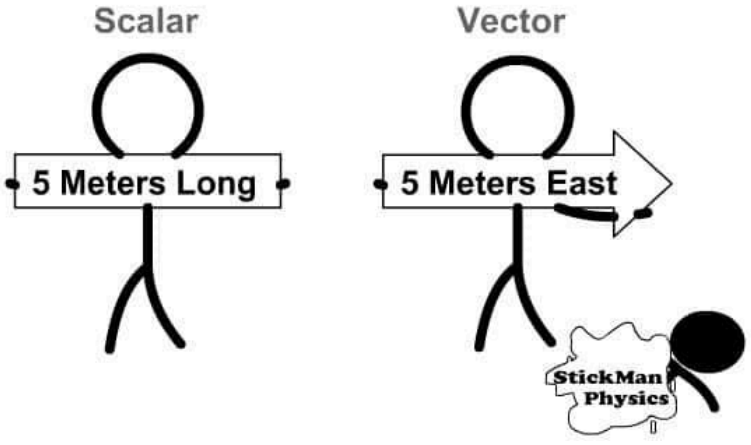
Vectors are utilised in day-to-day life to assist in the localization of people, places, and things. They are also used to describe things that are acting in response to an external force being applied to them. A quantity that possesses both a magnitude and a direction is known as a vector. The first, second, and third laws of Newton are all relationships between vectors that precisely describe the motion of bodies when they are subjected to the influence of an outside force. Newton's laws cover a wide range of phenomena and can be used to describe anything from a ball in free fall to a rocket on its way to the moon.

Q2. How is a vector different from a scalar?

Explanation - Every scientific explanation can be explained with the help of a number of physical quantities each expresses a special meaning and significance in that context. According to the definition, a ***physical quantity*** is the measurable and quantifiable physical property that carries unique information with it. Based on the dependency of direction, physical quantities can be classified into two categories — ***scalar and vector***.



Scalar and Vector Quantities are used to describe the motion of an object. ***Scalar Quantities*** are defined as physical quantities that have magnitude or size only. For example, distance, speed, mass, density, etc. However, ***vector quantities*** are those physical quantities that have both magnitude and direction like displacement, velocity, acceleration, force, etc. It should be noted that when a vector quantity changes its magnitude and direction also change similarly, when a scalar quantity changes, only its magnitude changes.



Difference between Scalar and Vector Quantity	
Scalar	**Vector**
1. Scalar quantities have magnitude or size only.	1. Vector quantities have both magnitude and direction.
2. It is known that every scalar exists in one dimension only.	2. Vector quantities can exist in one, two, or three-dimension.
3. Whenever there is a change in a scalar quantity, can correspond to a change in its magnitude also.	3. Any change in a vector quantity can correspond to change in either its magnitude or direction or both.
4. These quantities cannot be resolved into their components.	4. These quantities can be resolved into their components, using the sine or cosine of the adjacent angle.
5. Any mathematical process that involves more than two scalar quantities will only give scalars.	5. Mathematical operations on two or more vectors can provide either a scalar or a vector as a result. For instance, the dot product of two vectors only produces a scalar, whereas the cross product, sum, or subtraction of two vectors gives a vector.
Some examples of Scalar quantities are: ● Mass ● Speed ● Distance ● Time ● Area ● Volume	Some examples of Vector quantities are: ● Velocity ● Force ● Pressure ● Displacement ● Acceleration

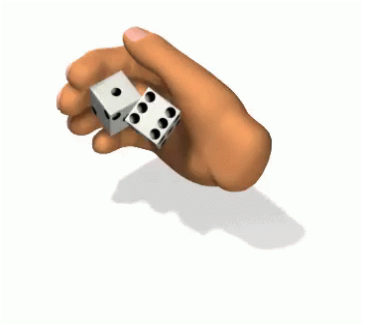
Q3. What is probability theory?

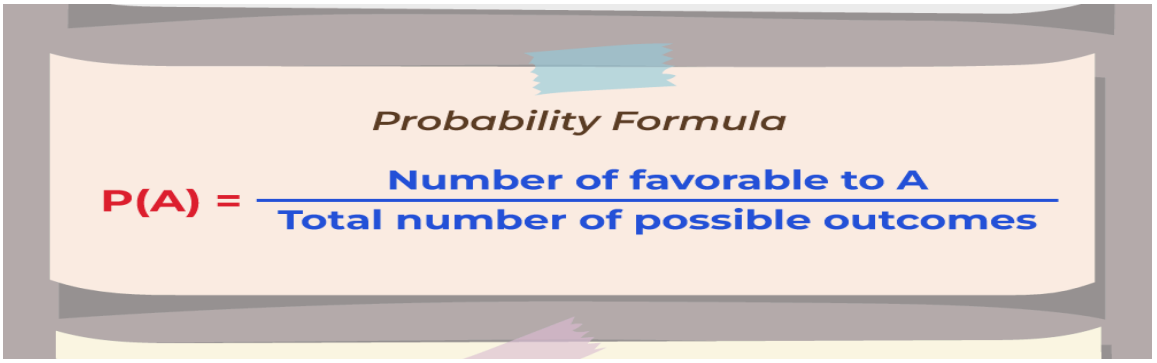
Introduction: Probability theory is a branch of mathematics that investigates the probabilities associated with a random phenomenon. A

random phenomenon can have several outcomes. Probability theory describes the chance of occurrence of a particular outcome by using certain formal concepts.

Definition: Probability theory makes the use of random variables and probability distributions to assess uncertain situations mathematically. In probability theory, the concept of probability is used to assign a numerical description to the likelihood of occurrence of an event. Probability can be defined as the number of favourable outcomes divided by the total number of possible outcomes of an event.

Below are the examples of probability theory:





Where,

A represents the event of interest.

P (A) is the probability of that event occurring.

Probability Theory Basics

There are some basic terminologies associated with probability theory that aid in the understanding of this field of mathematics.

Random Experiment

A random experiment, in probability theory, can be defined as a trial that is repeated multiple times in order to get a well-defined set of possible outcomes. Tossing a coin is an example of a random experiment.

Sample Space

Sample space can be defined as the set of all possible outcomes that result from conducting a random experiment. For example, the sample space of tossing a fair coin is {heads, tails}.

Event

Probability theory defines an event as a set of outcomes of an experiment that forms a subset of the sample space. The types of events are given as follows:

- ***Independent events***: Events that are not affected by other events are independent events.
- ***Dependent events***: Events that are affected by other events are known as dependent events.
- ***Mutually exclusive events***: Events that cannot take place at the same time are mutually exclusive events.
- ***Equally likely events***: Two or more events that have the same chance of occurring are known as equally likely events.
- ***Exhaustive events***: An exhaustive event is one that is equal to the sample space of an experiment.

Conclusion: Probability theory, a branch of mathematics concerned with the analysis of random phenomena. The outcome of a random event cannot be determined before it occurs, but it may be any one of several possible outcomes. The actual outcome is considered to be determined by chance.

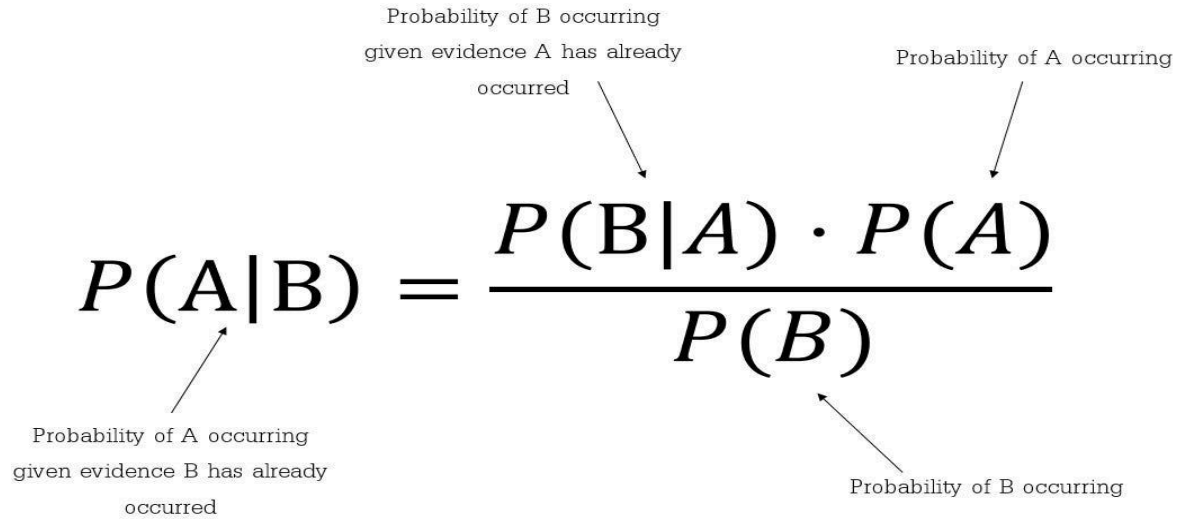
In short, probability is the study of uncertainty. It can be defined as a branch of mathematics dealing with randomness and uncertainty in concrete situations.

Q4.What is Bayes theorem, and how is it used?

Definition: In probability theory and statistics, **Bayes' theorem**, named after ***Thomas Bayes***, describes the probability of an event, based on prior knowledge of conditions that might be related to the event. **For example**, if the risk of developing health problems is known to increase with age, Bayes' theorem allows the risk to an individual of a known age to be assessed more accurately by conditioning it relative to their age, rather than simply assuming that the individual is typical of the population as a whole.

Statement of theorem:

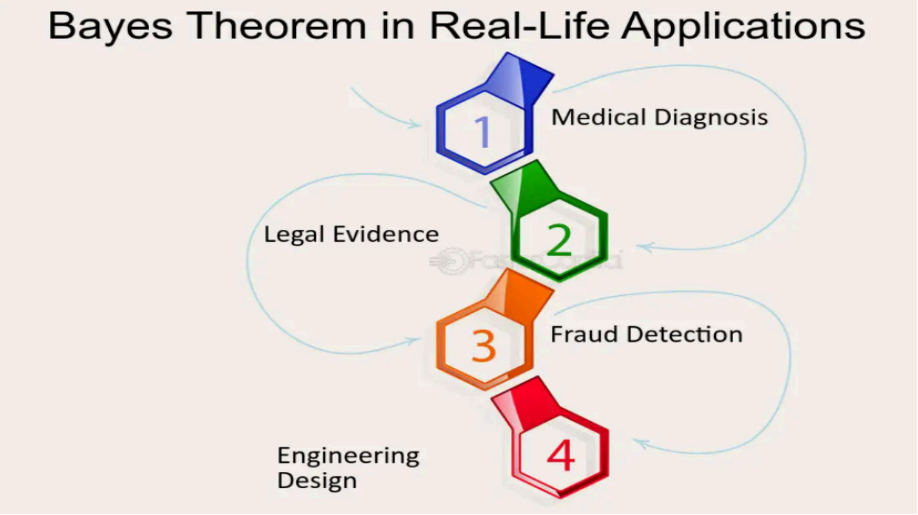
Bayes' theorem is stated mathematically as the following equation.



Bayes' Theorem thus gives the probability of an event based on new information that is or may be related to that event. The formula also can be used to determine how the probability of an event occurring may be affected by hypothetical new information, supposing the new information will turn out to be true.

Bayes Theorem in Real-Life Applications

Bayes' theorem is not just a mathematical concept or theoretical framework. It is a powerful tool that can be applied to real-life situations.



Medical Diagnosis: Bayes' theorem is used in medical diagnosis to determine the probability of a disease given the symptoms exhibited by a patient. For instance, if a patient exhibits symptoms such as coughing, fever, and shortness of breath, the doctor can use Bayes' theorem to calculate the probability of the patient having a particular disease, such as pneumonia or COVID-19, based on the prevalence of the disease in the population and the accuracy of the test used to detect the disease.

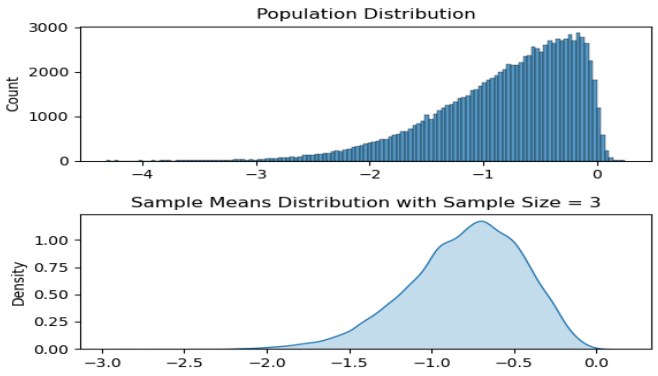
Fraud Detection: Bayes' theorem is used in fraud detection to identify suspicious transactions or activities. For instance, if a bank receives a transaction that is outside the normal pattern of a customer's behaviour, Bayes' theorem can be used to calculate the probability of the transaction being fraudulent based on the prevalence of fraud in the population and the likelihood of the customer's behaviour changing.

Bayes Theorem is also used widely in machine learning, where it is a simple, effective way to predict classes with precision and accuracy. The Bayesian method of calculating ***conditional probabilities*** is used in machine learning applications that involve classification tasks.

Hence, we can say that Machine Learning is highly dependent on Bayes theorem.

Q5.What is the central limit theorem, and how is it used?

Definition: Central Limit Theorem, also known as the CLT, is a crucial pillar of statistics and machine learning. The central limit theorem states that if you take sufficiently large samples from a population, the samples' means will be normally distributed, even if the population isn't normally distributed.



The sample size of 30 is considered sufficient to see the effect of the CLT. If the population distribution is closer to the normal distribution, you will need fewer samples to demonstrate the central limit theorem. On the other hand, if the population distribution is highly skewed, you will need a large number of samples to understand the CLT.

Understanding CLT with Example:

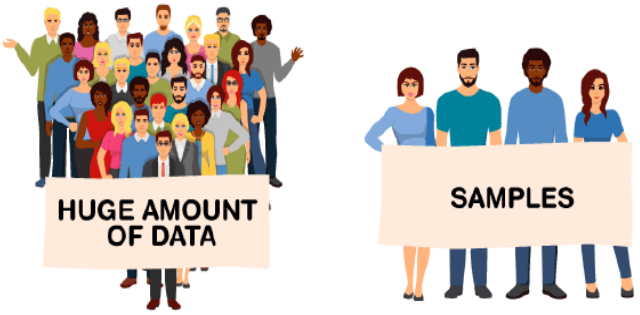
Let’s understand the central limit theorem with the help of an example..

Consider that there are 15 Cohort in the data science department of Alma better, and each cohort hosts around 100 students. Our task is to calculate the average weight of students in data science department. Sounds simple, right?
The approach I get from aspiring data scientists is to simply calculate the average:

First, measure the weights of all the students in the data science department. Add all the weights.

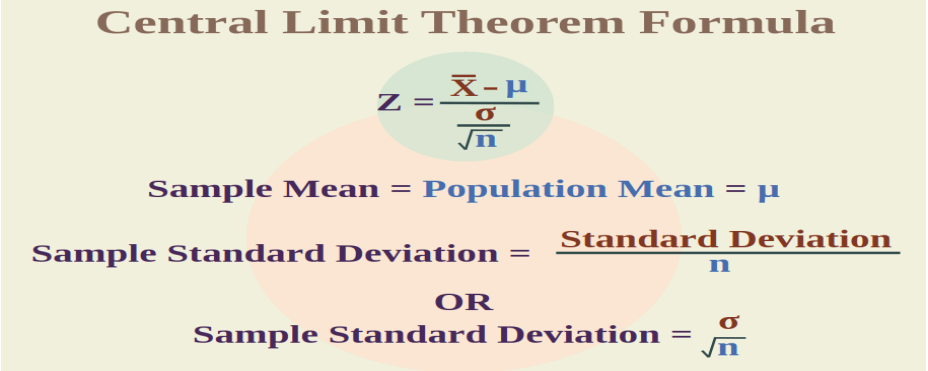
- Finally, divide the total sum of weights by the total number of students to get the average.

But what if the size of the data is humongous? Does this approach make sense? Not really – measuring the weight of all the students will be a very tiresome and long process. So, what can we do instead? Let's look at an alternate approach.
First, draw groups of students at random from the class. We will call this a sample. We'll draw multiple samples, each consisting of 30 students.



Now, calculate the individual mean of these samples.
Then, calculate the mean of these sample means.
This value will give us the approximate mean weight of the students in the science department.
Additionally, the histogram of the sample mean weights of students will resemble a bell curve (or normal distribution).

Central Limit Theorem Formula



Conditions of the Central Limit Theorem

1. The sample size is **sufficiently large**. This condition is usually met if the size of the sample is $n \geq 30$.
2. The samples are **independent and identically distributed, i.e., random variables**. The sampling should be random.
3. The population's distribution has a **finite variance**. The central limit theorem doesn't apply to distributions with infinite variance.

Uses of CLT

The central limit theorem is useful when analysing large data sets because it allows one to assume that the sampling distribution of the mean will be normally-distributed in most cases. This allows for easier statistical analysis and inference.

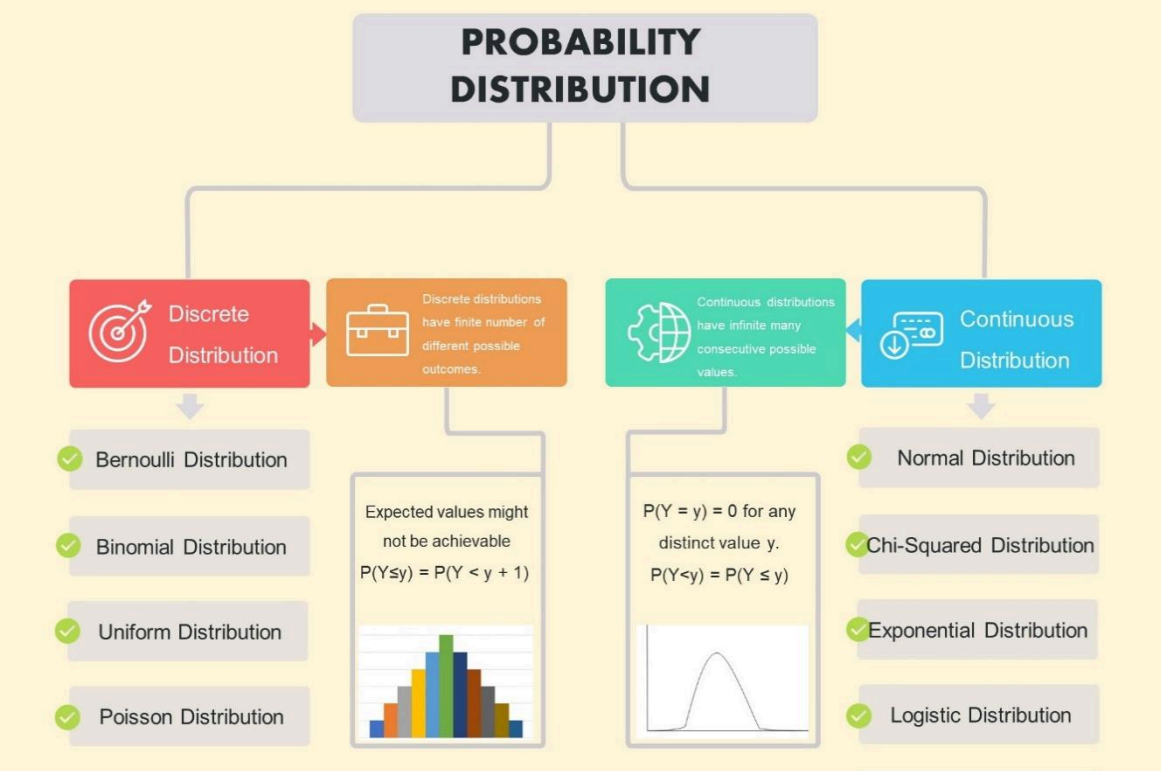
We use the CLT in various census fields to calculate various population details, such as family income, electricity consumption, individual salaries, and so on.

In quality control, the CLT is used to determine whether a production process is operating within acceptable limits. In election polling, the CLT is used to estimate the proportion of voters who support a particular candidate.

Q6.What is the difference between discrete and continuous probability distributions?

A probability distribution is a statistical function that describes all the possible values and probabilities for a random variable within a given range. This range will be bound by the minimum and maximum possible values, but where the possible value would be plotted on the probability distribution will be determined by a number of factors. The mean (average), standard deviation, skewness, and kurtosis of the distribution are among these factors.

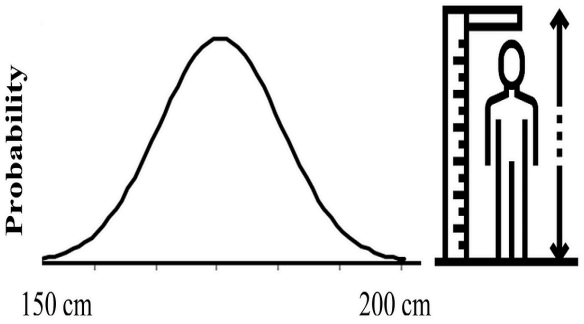
Types of Probability Distribution



Discrete probability distribution: A discrete probability distribution gives the likelihood of occurrence of each possible value of a discrete random variable. The number of spoiled apples out of 6 in your refrigerator can be an example of a discrete probability distribution.

Continuous Probability Distributions: A continuous distribution describes the probabilities of a continuous random variable's possible values. A continuous random variable has an infinite and uncountable set of possible values (known as the range). The mapping of time can be considered as an example of the continuous probability distribution. It can be from 1 second to 1 billion seconds, and so on.

Difference between

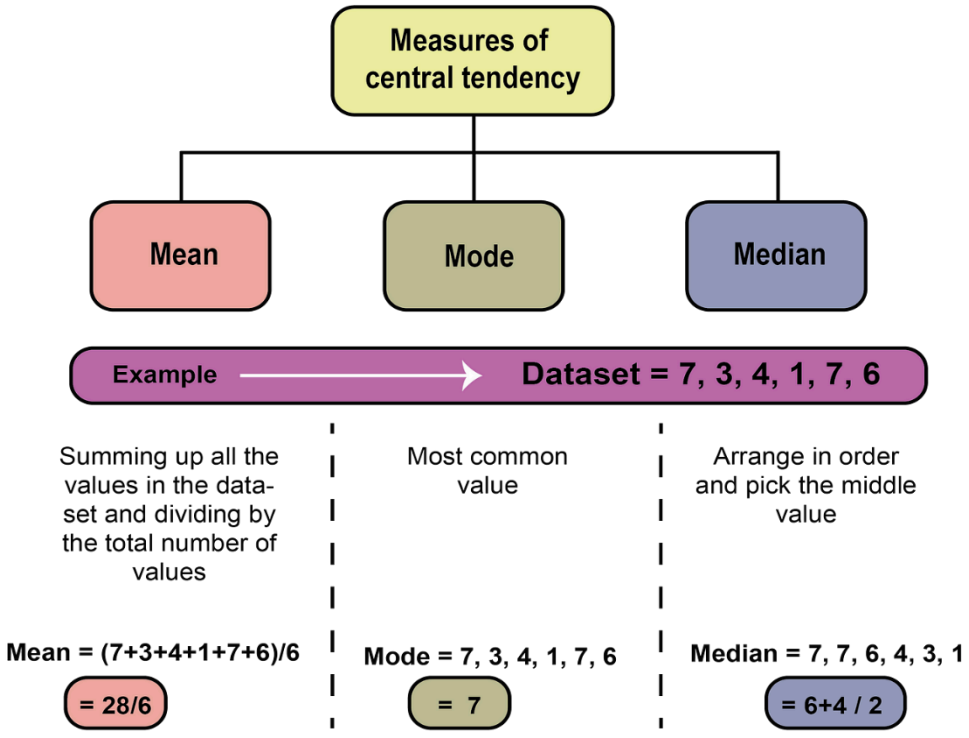
Discrete probability distribution	***Continuous probability distribution***
Countable set of distinct values	Any value within some interval (say 1 to 2)
Discrete data is counted	Continuous data is measured
Can take only integer values. Never include fractions or decimals.	Can take values including fractions and decimals.
Discrete data can only take certain values. Ex: Number of students in a class (you cannot have 56.5 students)	Continuous data can take any value, including decimal points (within a range) Ex: A person's height (167.54 cm) could be any value (within a range of human heights: 40 to 270 centimetres)
Examples: • Number of children in a family • Number of defective bulbs in a box of 10 • Number of ants born tomorrow • Number of classes missed last week (0,1,2..) • Toss of a coin • Number of heads in 4 flips of a coin (possible outcomes: 0,1,2,3,4) • Number of patients in hospital	Examples: • Amount of sugar in a coffee • Amount of rain in a day • Time to finish a test • Percentage of marks obtained by a student • Length of a chord of a circle (any number of decimal places) • Height of individuals • Hours spent exercising last week • Time required to finish a test
	
Probability Mass Function	Probability Density Function
Count,Sum,Proportion	Integration
$P(X = x) = f(x)$	$P(X=x) = \int f(x).dx$
CMF,PMF = Sum, Difference	CDF,PDF = Integrate, Differentiate

Q7.What are some common measures of central tendency, and how are they calculated?

Explanation: Measures of central tendency are the values that describe a data set by identifying the central position of the data. It is defined as the statistical measure that can be used to represent the entire distribution or a dataset using a single value called a measure of central tendency. Any of the measures of central tendency provides an accurate description of the entire data in the distribution.

There are generally three measures of central tendency, commonly used in statistics- ***mean, median, and mode.*** Mean is the most common measure of central tendency used to describe a data set.

Let's understand diagrammatic representation of measures of central tendency:



Mean- Sum of all observations divided by the total number of observations.

Mode- The most frequently occurring value in a data set.

Median- The middle or central value in an ordered set.

Mean as a Measure of Central Tendency

We generally denote the mean of a given data-set by $\bar{x}$, pronounced “x bar”. The formula to calculate the mean for ungrouped data to represent it as the measure is given as,

For a set of observations: **Mean = Sum of the terms / Number of terms**

For a set of grouped data: Mean, $\bar{x} = \Sigma fx/\Sigma f$

where,

- $\bar{x}$ = the mean value of the set of given data.
- f = frequency of each class
- x = mid-interval value of each class

Median as a Measure of Central Tendency

The major advantage of using the median as a central tendency is that it is less affected by outliers and skewed data. We can calculate the median for different types of data, grouped data, or ungrouped data using the median formula.For ungrouped data: For odd number of observations, Median = [(n + 1)/2]th term. For even number of observations, Median = [(n/2)th term + ((n/2) + 1)th term]/2

For grouped data: Median = $l + [((n/2) - c)/f] \times h$

where,

l = Lower limit of the median class

c = Cumulative frequency

h = Class size

n = Number of observations

Median class = Class where n/2 lies

Mode as a Measure of Central Tendency

Mode is defined as the value which appears most often in the given data, i.e. the observation with the highest frequency is called the mode of data.

Mode for ungrouped data: Most recurring observation in the data set.

Mode for grouped data: $\mathbf{L + h \frac{(fm-f1)}{(fm-f1) + (fm-f2)}}$

Where,

- L is the lower limit of the modal class
- h is the size of the class interval
- fm is the frequency of the modal class
- f1 is the frequency of the class preceding the modal class
- f2 is the frequency of the class succeeding the modal class

Q8: What is the purpose of using percentiles and quartiles in data summarization?

Explanation: Percentiles and quartiles are important measures of relative standing or position of a value in a data set.

Percentiles divide a dataset into 100 equal parts, each representing 1% of the data. For example, the 75th percentile is the value below which 75% of the data falls. It is commonly expressed as the percentage of values in a set of data scores that fall below a given value.

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

Percentiles and quartiles can be calculated using various methods, such as interpolation, nearest rank, and the inverse of the cumulative distribution function. The most commonly used method for calculating percentiles and quartiles is the interpolation method. This method estimates the percentile value by interpolating between the two nearest values in the dataset.

In Python, you can use the NumPy library to calculate percentiles and quartiles. The percentile () function can be used to calculate any percentile value, and the quantile () function can be used to calculate quartiles.

Example code for calculating percentiles:

```
import numpy as np

data = [3, 1, 4, 2, 5, 7, 6, 8, 9, 10]
p75 = np.percentile(data, 75)
print("75th percentile value is:", p75)
```

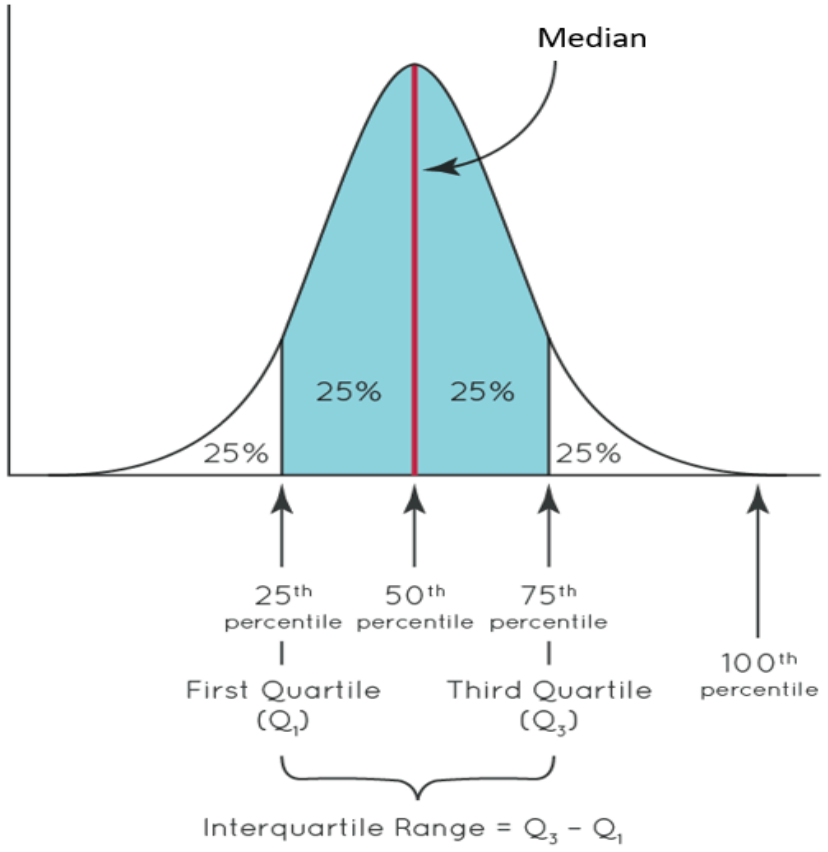
```
75th percentile value is: 7.75
```

Example code for calculating quartiles:

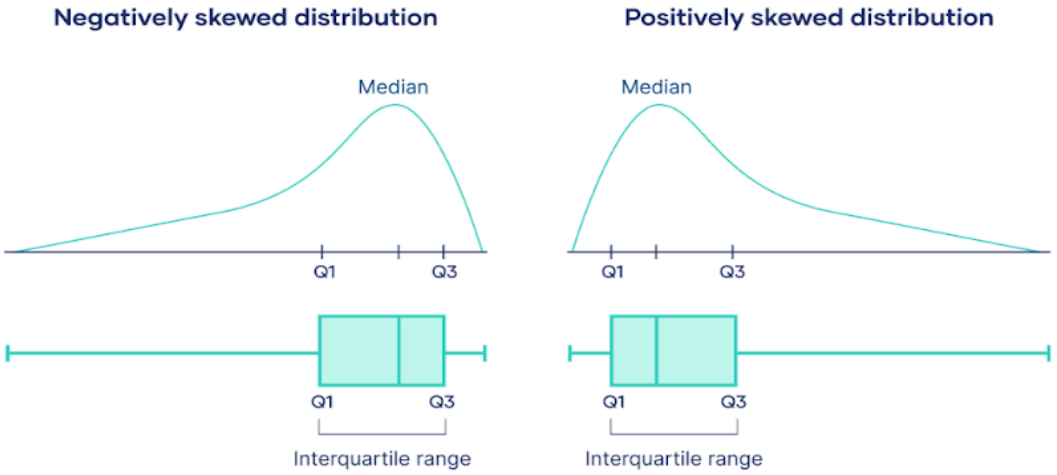
```
import numpy as np

data = [3, 1, 4, 2, 5, 7, 6, 8, 9, 10]
q1 = np.quantile(data, 0.25)
q2 = np.quantile(data, 0.5)
q3 = np.quantile(data, 0.75)

print("Q1 value is:", q1)
print("Q2 value is:", q2)
print("Q3 value is:", q3)
```

```
Q1 value is: 3.25
Q2 value is: 5.5
Q3 value is: 7.75
```

Diagrammatic representation of Quartiles and Percentiles



Purpose: Quartiles are helpful for understanding an observation in the context of the rest of a sample or population. By comparing the observation to the quartiles, we can determine whether the observation is in the bottom 25%, middle 50%, or top 25%. The second quartile, better known as the **median**, is a measure of central tendency. This middle number is a good measure of the average or most central value of the data, especially for skewed distributions or distributions with outliers.

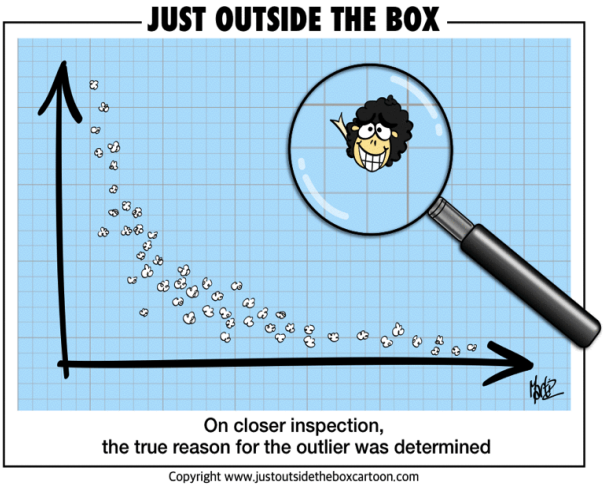


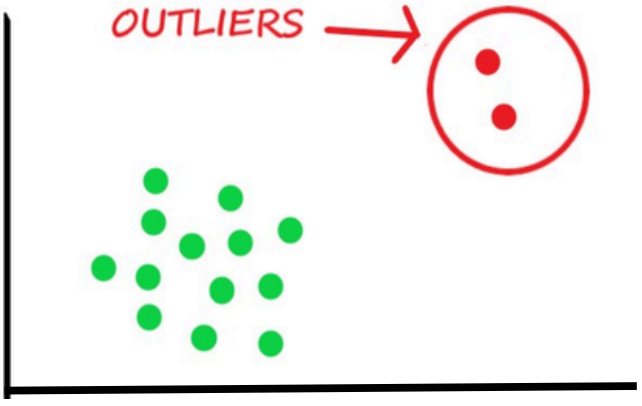
The distance between the first and third quartiles—the interquartile range (IQR)—is a measure of variability. It indicates the spread of the middle 50% of the data. The distance between quartiles can give a hint about whether a distribution is skewed or symmetrical. It's easiest to use a boxplot to look at the distances between quartiles.

Q9.How do you detect and treat outliers in a dataset?

We all have heard of the idiom '***odd one out*** which means something unusual in comparison to the others in a group. Similarly, an Outlier is an observation in a given dataset that lies far from the rest of the observations. That means an outlier is vastly larger or smaller than the remaining values in the set.

An outlier is a data point that differs significantly from other observations. An outlier may be due to a variability in the measurement, an indication of novel data, or it may be the result of experimental error; the latter are sometimes excluded from the data set. An outlier can be an indication of exciting possibility, but can also cause serious problems in statistical analyses.





Causes of outliers:

- Human errors, e.g. data entry errors
- Instrument errors, e.g. measurement errors
- Data processing errors, e.g. data manipulation
- Sampling errors, e.g. extracting data from wrong sources
- Natural novelties in data: The outliers that are not caused due to any error are called Natural Outliers. For example, in a class of 50 students, 45 students perform average in a test while 3 students perform excellently it and 2 students perform poorly in the test. Now the

students who performed excellently and the students who performed poorly are outliers, but they are not caused due to an error.

Outliers are the extreme deviated values in data that can cause variances in results and can impact our analysis outcomes. There are many causes of outliers in a data set such as sampling errors and measurement errors. Before dealing with outliers we also need to detect the outliers, and this can be done via methods like box plot, using the Z-scores and using the Inter Quantile Range (IQR).

Detecting Outliers

If our dataset is small, we can detect the outlier by just looking at the dataset. But what if we have a huge dataset, how do we identify the outliers then? We need to use visualization and mathematical techniques.

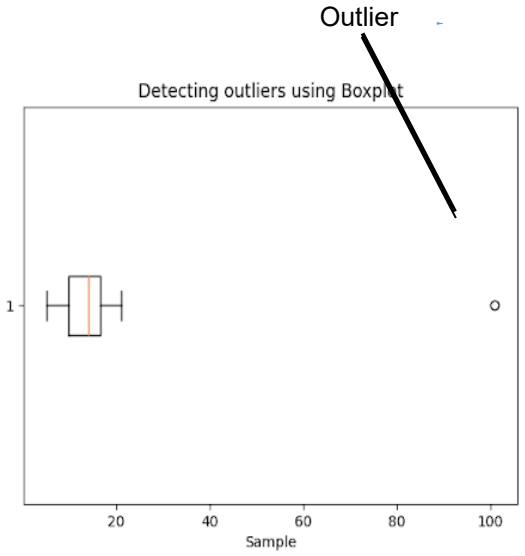
Below are some of the techniques of detecting outliers

- Boxplots
- Z-score
- Inter Quantile Range (IQR)

Detecting Outliers Using Boxplot

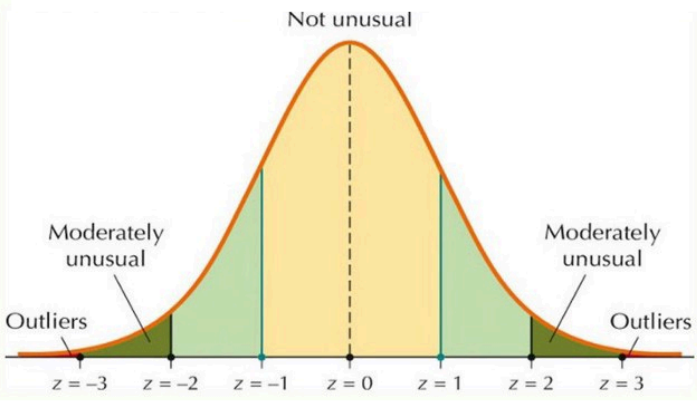
Python code for boxplot is:

```
1  import matplotlib.pyplot as plt
2
3  sample= [15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10, 9]
4  plt.boxplot(sample, vert=False)
5  plt.title("Detecting outliers using Boxplot")
6  plt.xlabel('Sample')
7  plt.show()
```



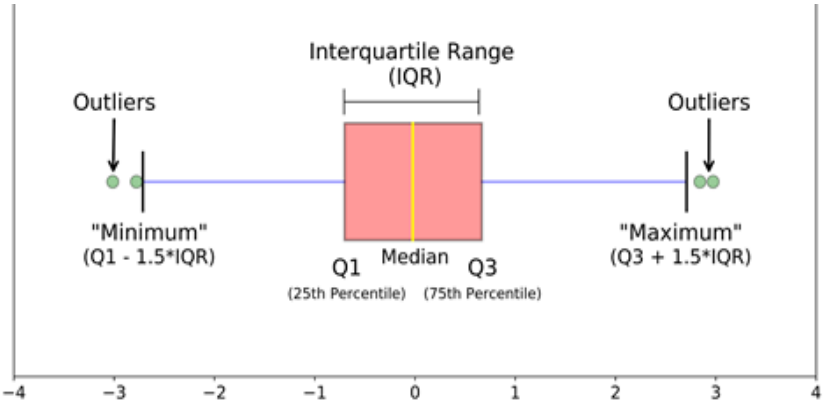
Detecting Outliers using the Z-scores

Criteria: any data point whose Z-score falls out of 3rd standard deviation is an outlier.



- Loop through all the data points and compute the Z-score using the formula (Xi-mean)/std.
- Define a threshold value of 3 and mark the data points whose absolute value of Z-score is greater than the threshold as outliers.

Detecting Outliers using the Inter Quantile Range (IQR)



Criteria: data points that lie 1.5 times of IQR above Q3 and below Q1 are outliers.

Steps

- Sort the dataset in ascending order
- calculate the 1st and 3rd quartiles(Q1, Q3)
- compute IQR=Q3-Q1
- compute lower bound = (Q1–1.5*IQR), upper bound = (Q3+1.5*IQR)

- loop through the values of the dataset and check for those who fall below the lower bound and above the upper bound and mark them as outliers

Treatment of outliers

Methods of dealing with outliers, apart from removing them from the dataset:
1) Reducing the weights of outliers (trimming weight)

2) Changing the values of outliers (Winsorisation, trimming, imputation): As the mean value is highly influenced by the outliers, it is advised to replace the outliers with the median value.

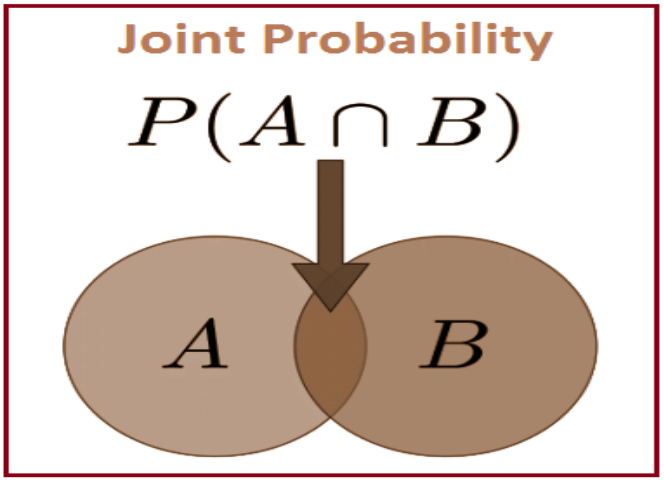
3) Quantile Based Flooring and Capping: In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value. The data points that are lesser than the 10th percentile are replaced with the 10th percentile value and the data points that are greater than the 90th percentile are replaced with 90th percentile value.

Outliers, once understood and managed, become valuable sources of information, ultimately contributing to more informed and reliable decision-making processes.

We should not just drop the outliers from our analysis since in certain cases outliers can give valuable information about our processes. There are lots of ways to deal with outliers in data and there is no quick fix or magic to handle them - in most cases human expertise and experience comes into play to decide how to best handle outliers in our data. ***Outliers can be genuine observations… be gentle to the data and document each and every step of the data processing.***

Q10. What is a joint probability distribution?

Joint probability is the probability of two events happening together, and their joint probability distribution is the corresponding probability distribution on all possible outcomes of those events. A **joint probability distribution** simply describes the probability that a given individual takes on two specific values for the variables. The word "joint" comes from the fact that we're interested in the probability of two things happening at once.



Let's take an example to understand joint probability distribution:

Out of the 100 total individuals there were 13 who were male *and* chose baseball as their favourite sport. Thus, we would say the joint probability that a given individual is male *and* chooses baseball as their favourite sport is 13/100 = **0.13** or **13%**.

Written in mathematical notation:

We can use this process to calculate the entire joint probability distribution:

- P(Gender = Male, Sport = Baseball) = 13/100 = **0.13**
- P(Gender = Male, Sport = Basketball) = 15/100 = **0.15**
- P(Gender = Male, Sport = Football) = 20/100 = **0.20**
- P(Gender = Female, Sport = Baseball) = 23/100 = **0.23**
- P(Gender = Female, Sport = Basketball) = 16/100 = **0.16**
- P(Gender = Female, Sport = Football) = 13/100 = **0.13**

Notice that the sum of the probabilities is equal to 1, or 100%.

A joint probability distribution shows a probability distribution for two (or more) random variables. Instead of events being labelled A and B, the norm is

to use X and Y. The formal definition is*: f(x, y) = P(X = x, Y = y).* ***The whole point of the joint distribution is to look for a relationship between two variables.***

Q11. How do you calculate the joint probability distribution?

Joint Probability Distribution is used to describe general situations where several random variables like *X* and *Y* are observed which is similar to experimental probability. The joint probability mass function or the joint density is used to compute probabilities involving such variables as *X* and *Y*.

***Example of Joint Probability Distribution*:** We have a box of ten balls in which four balls are white, three balls are red, and three balls are black. Here the number of red balls selected is *X* and the number of white balls selected is *Y*. If we select five balls out of the box without replacement and count the number of white and red balls in the sample, then we can find probabilities of any event involving *X* and *Y*, using the Joint Probability Distribution table. Using the Joint Probability Distribution table we can find the probability that one samples the same number of red and white balls or the probability one samples more red balls than white balls and so on.

Let joint probability distribution shows a probability distribution for two (or more) random variables.

The formal definition of a joint probability distribution can be written as:

f(x, y)=P(X=x, Y=y)

We use the Joint Probability Distribution to look for a relationship between two variables.

Example of Joint Probability Distribution for a relationship between two variables: We have a box of ten balls in which four are white, three are black, and three are red. One has to select five balls out of the box without replacement and count the number of white and red balls in the sample. What is the probability one observes two white and two red balls in the sample?

Here, the total number of outcomes is $^{10}C5=25210$

Next, one thinks about the number of ways of selecting two white and two red balls. One does this in steps – first select the white balls, then select the red balls, and then select the one remaining black ball. Note that five balls are selected, so exactly one of the balls must be black. Since the box has four white

$$P(X=2,\ Y=2)=\frac{{}^{4}C_{2}\times{}^{3}C_{2}\times{}^{3}C_{1}}{{}^{10}C_{5}}=\frac{54}{252},$$

balls, the number of ways of choosing two white is $^4C2$. Of the three red balls, one wants to choose two – the number of ways of doing that is $^3C2$. Last, the number of ways of choosing the remaining one black ball is $^3C1$. So the total number of ways of choosing two white, two red, and one black ball is the product,

Where *X*=number of red balls selected, *Y*=number of white balls selected.

Suppose this calculation is done for every possible pair of values of *X* and *Y.* These possibilities can be tabulated as shown below. This table is known as the **Joint Probability Distribution Table for *X* and *Y***.

X=number of red balls selected→ Y=number of white balls selected↓	0	1	2	3	4
0	0	0	6/252	12/252	3/252
1	0	12/252	54/252	36/252	3/252
2	3/252	36/252	54/252	12/252	0
3	3/252	12/252	6/252	0	0

This table is called the joint probability mass function (pmf) f(x, y) of (X, Y). As for any probability distribution, one requires that each of the probability values is nonnegative and the sum of the probabilities over all values of X and Y is one. That is, the function f(x, y) satisfies two properties as mentioned below.

1. $f(x, y) \geq 0, \forall x, y$
2. $\Sigma x, yf(x, y)=1$

Joint probability provides insights into the relationship between events. If the joint probability is high, it suggests a strong association between the events, indicating that they are more likely to occur together. Conversely, a low joint probability implies a weak association or independence between the events. Joint probability finds applications in diverse fields. For instance, in medical research, joint probability is used to assess the likelihood of multiple risk factors occurring simultaneously. In finance, it helps determine the joint probabilities of different assets' returns. Additionally, it plays a vital role in decision-making under uncertainty and modelling real-world scenarios.

Q12. What is the difference between a joint probability distribution and a marginal probability distribution?

The concepts of probability are fundamental to machine learning and data science. While it is easy to understand and model a single random variable, in practice, we usually have many random variables that may interact with each other.

A joint distribution is a probability distribution that describes the probability of two or more random variables having specific values at the same time.

A marginal distribution is a probability distribution that describes the probability of one random variable having a specific value, regardless of the value of any other random variables. The marginal distribution is obtained by summing or integrating the joint distribution over the values of the other random variables. For example, the marginal distribution of a random variable X is represented by the function $P(X = x) = \sum_y P(X = x, Y = y)$, where x is a specific value of the random variable X and y is a variable representing the values of another random variable Y.

The marginal distribution is obtained by summing (or integrating, in the case of continuous variables) the joint probability distribution over the variables not of interest.

Marginal distributions can be calculated for both discrete and continuous variables.

Marginal distribution is useful in simplifying the analysis of complex problems that involve multiple variables.

The marginal distribution of a single variable can be obtained by summing (or integrating) the joint distribution over all possible values of the other variables.

The marginal distribution of multiple variables can be obtained by summing (or integrating) the joint distribution over all possible values of the variables not of interest.

Marginal distributions are important in statistics, as they allow us to study the behaviour of individual variables in a multivariate distribution.

In Python, the marginal distribution can be calculated using the numpy.sum () function for discrete variables and scipy.integrate.simps() function for continuous variables. Here's an example:

```
import numpy as np
import scipy.integrate as spi

# Define the joint probability density function
def joint_pdf(x, y):
    return x*y*np.exp(-x*y)

# Define the marginal probability density function for x
def marginal_pdf_x(x):
    return spi.simps(joint_pdf(x, y_vals), y_vals)

# Define the marginal probability density function for y
def marginal_pdf_y(y):
    return spi.simps(joint_pdf(x_vals, y), x_vals)

# Define the range of values for x and y
x_vals = np.linspace(0, 5, 50)
y_vals = np.linspace(0, 5, 50)

# Calculate the marginal PDFs for x and y
marginal_x = np.array([marginal_pdf_x(x) for x in x_vals])
marginal_y = np.array([marginal_pdf_y(y) for y in y_vals])
```

The resulting arrays marginal_x and marginal_y contain the marginal probability density functions for the variables x and y, respectively.

Q13. What is the covariance of a joint probability distribution?

The covariance of a joint probability distribution measures the degree to which two random variables change together. In the context of probability distributions, covariance is a measure of how much two random variables vary together. For two random variables X and Y with a joint probability distribution P(X, Y), the covariance (denoted as Cov(X, Y)) is calculated as follows:

$$Cov(X,Y)=E[(X-\mu X)(Y-\mu Y)]$$

Where,

- E denotes the expected value,
- μX is the mean of the random variable X,
- μY is the mean of the random variable Y.

Alternatively, the covariance can be expressed as:

$$Cov(X,Y)=E[XY]-E[X]E[Y]$$

The sign of the covariance indicates the direction of the linear relationship between the two variables:

- Positive covariance suggests that as one variable increases, the other tends to increase as well.
- Negative covariance suggests that as one variable increases, the other tends to decrease.

Here is an example of how to calculate covariance in python:

```
import numpy as np

# Sample data
X = np.array([1, 2, 3, 4, 5])
Y = np.array([5, 4, 3, 2, 1])
# Calculate covariance
 covariance = np.cov(X, Y)[0][1]

print("Covariance:", covariance)
```

```
Covariance: -2.5
```

However, the magnitude of the covariance is not easily interpretable in terms of the strength of the relationship, as it depends on the scales of the variables. Therefore, the correlation coefficient is often used to standardize the covariance and provide a more interpretable measure of the strength and direction of the linear relationship between two variables.

Q14. What is the relationship between the correlation coefficient and the covariance of a joint probability distribution?

The correlation coefficient and the covariance are both measures of the relationship between two variables in a joint probability distribution.

Covariance (denoted as Cov(X, Y)) measures how much two random variables X and Y change together. Specifically, it measures the average product of the deviations of each variable from their respective means.

The formula for covariance is given by:

$$\mathrm{Cov}(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Where

- Xi and Yi are individual data points,
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y,
- N is the number of data points.

The correlation coefficient (denoted as r) is a standardized measure of the strength and direction of a linear relationship between two variables. It is calculated by dividing the covariance of the variables by the product of their standard deviations. The formula for the correlation coefficient is given by:

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Here, σX and σY are the standard deviations of X and Y, respectively.

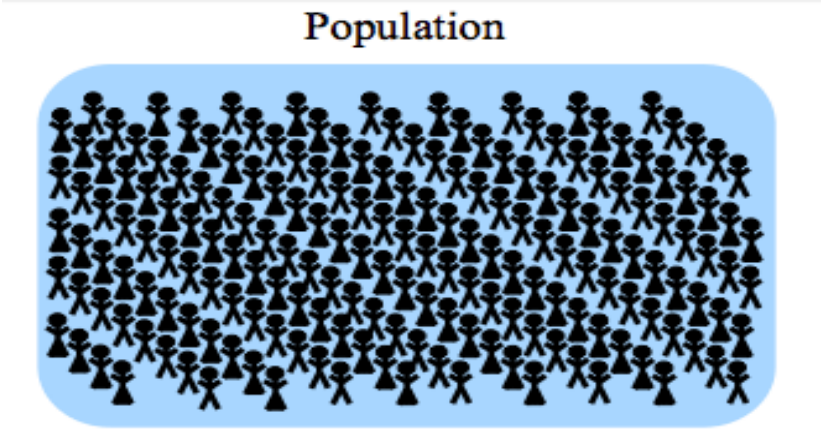
The correlation coefficient is the normalized version of covariance. While covariance measures the extent to which two variables change together, the correlation coefficient additionally scales this measure by dividing it by the product of the standard deviations of the two variables.

CORRELATION	COVARIANCE
There is said to be correlation between two, when change in one results in change in another.	Covariance talks about the direction of the relationship between the two variables (positive or negative)
CORRELATION	**COVARIANCE**
Measures the strength of the variables under comparison	Measures the extent of change in one with regards to change in another.
Correlation is a scaled down version of covariance.	Covariance is considered as a part of correlation.
Value here lies between -1 and +1.	Value here lies between -infinity to +infinity
Correlation is a unit-free measure	Covariance value is the product of the units of the variables.
There would be no change in correlation due to scale.	Any change in scale affects covariance.

In summary, Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables. Correlation values are standardized. .the correlation coefficient is the covariance divided by the product of the standard deviations. By dividing by the standard deviations, the correlation coefficient is normalized and falls within the range of -1 to 1, making it easier to interpret in terms of the strength and direction of the relationship between the variables. If *r* is close to 1, it indicates a strong positive linear relationship, while if it's close to -1, it indicates a strong negative linear relationship. If *r* is close to 0, it suggests a weak or no linear relationship.

Q15. What is sampling in statistics, and why is it important?

Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. The population refers to the entire group of individuals, items, or data points that share a common set of characteristics, while the sample is a representative subset of that population. Sampling allows researchers to conduct studies about a large group by using a small portion of the population. The method of sampling depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling. Sampling is commonly done in statistics, psychology, and the financial industry.



It can be difficult for researchers to conduct accurate studies on large populations. In some cases, it can be impossible to study every individual in the group. That's why they often choose a small portion to represent the entire group. This is called a sample. Samples allow researchers to use characteristics of the small group to make estimates of the larger population.

The chosen sample should be a fair representation of the entire population. When taking a sample from a larger population, it is important to consider how the sample is chosen. To get a representative sample, it must be drawn randomly and encompass the whole population. For example, a lottery system could be used to determine the average age of students in a university by sampling 10% of the student body.

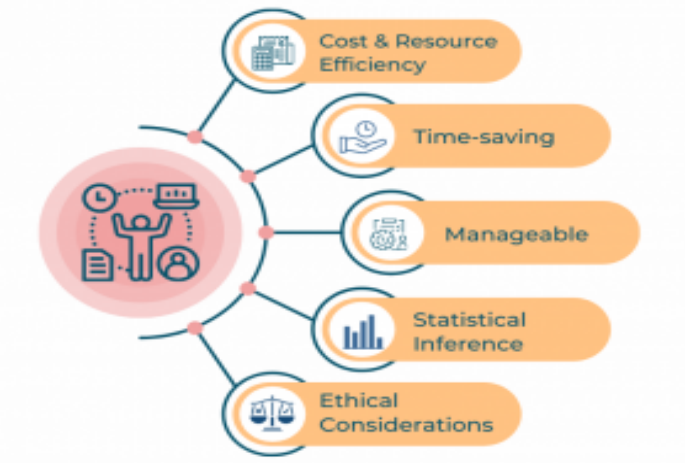
Importance of Sampling

Sampling is of paramount importance in statistics for several reasons:

Cost efficiency: Studying an entire population can be impractical or cost-prohibitive. Sampling allows researchers to gather information from a subset of the population, reducing the time and resources required.

Time efficiency: Analysing a sample is often quicker than analysing an entire population. This is particularly relevant when timely decisions or results are needed.

Feasibility: In cases where the population is vast, dispersed, or difficult to access, sampling provides a practical way to collect data without the challenges associated with studying the entire population.



Statistical inference: Sampling is fundamental to statistical inference, where conclusions about a population are drawn from the analysis of a representative sample. This allows researchers to make predictions, test hypotheses, and estimate population parameters.

Practicality: In some situations, it's simply not possible to study an entire population. Sampling allows statisticians to work with manageable data sets while still drawing meaningful conclusions.

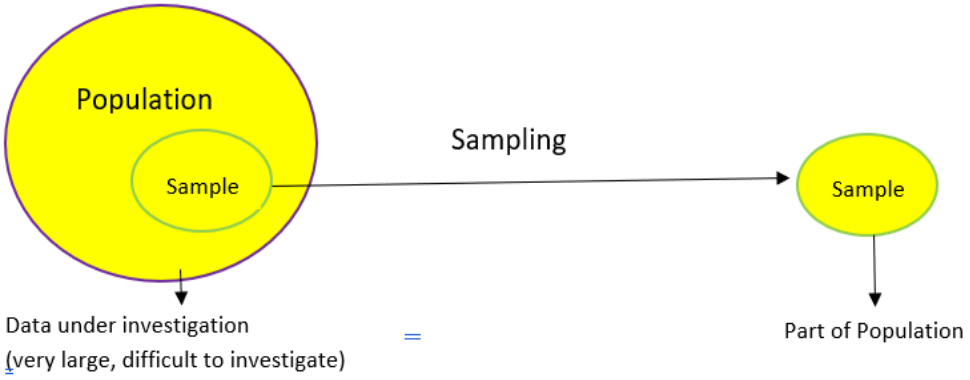
Generalizability: If a sample is carefully selected and representative of the population, the findings from the sample can often be generalized to the entire population. This is the basis for inferential statistics.

Reduced variability: While a sample may not perfectly represent the population, it can provide a good estimate. Sampling helps reduce variability, and statistical methods can be used to quantify the level of uncertainty in the estimates.

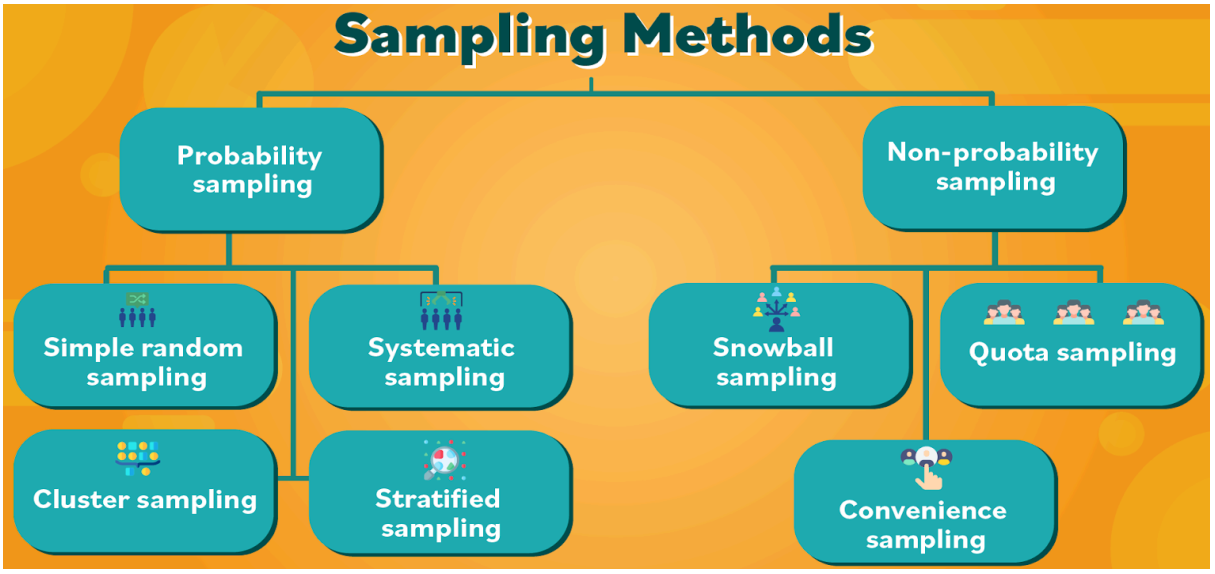
Resource conservation: Limited resources, such as manpower and financial resources, can be efficiently allocated when working with a sample rather than the entire population.

Q16. What are the different sampling methods commonly used in statistical inference?

Sampling is done to draw conclusions about populations from samples, and it enables us to determine a population's characteristics by directly observing only a portion (or sample) of the population.



Different sampling methods



- **Probability Sampling:** In probability sampling, every element of the population has an equal chance of being selected. Probability sampling

gives us the best chance to create a sample that is truly representative of the population.

- **Non-Probability Sampling:** In non-probability sampling, all elements do not have an equal chance of being selected. Consequently, there is a significant risk of ending up with a non-representative sample which does not produce generalizable results.

Types of Probability sampling

Simple Random Sampling

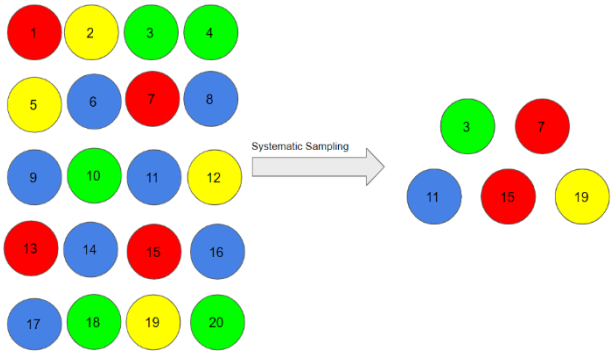
This is a type of sampling technique you must have come across at some point. Here, every individual is chosen entirely by chance and each member of the population has an equal chance of being selected. One big advantage of this technique is that it is the most direct method of probability sampling. But it comes with a caveat – it may not select enough individuals with our characteristics of interest.

Systematic sampling

In this type of sampling, the first individual is selected randomly and others are selected using a fixed 'sampling interval'.

Suppose, we began with person number 3, and we want a sample size of 5. So, the next individual that we will select would be at an interval of (20/5) = 4

from the 3rd person, i.e. 7 (3+4) and so on 3, 3+4=7, 7+4=11, 11+4=15, 15+4=19 = 3, 7, 11, 15, 19.

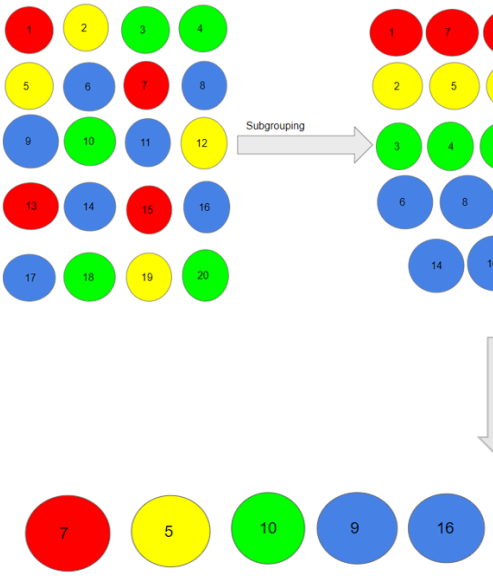


Systematic sampling is more convenient than simple random sampling. However, it might also lead to bias if there is an underlying pattern in which we

are selecting items from the population (though the chances of that happening are quite rare).

Stratified Sampling

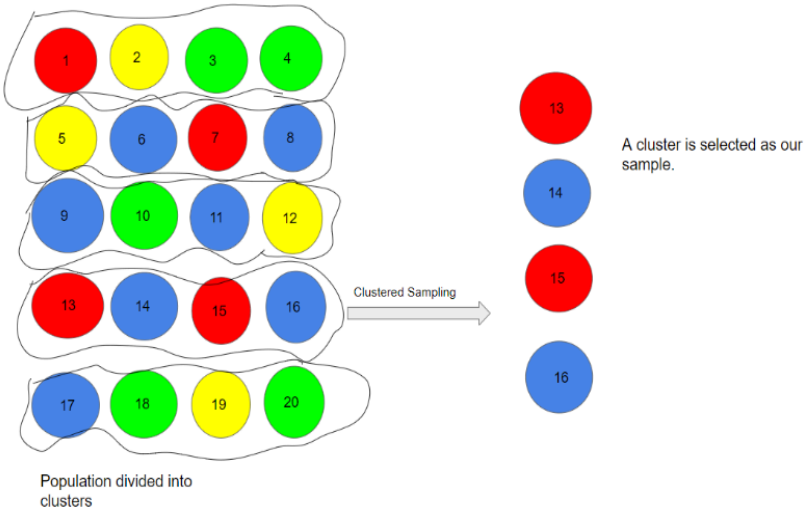
In this type of sampling, we divide the population into subgroups (called strata) based on different traits like gender, category, etc. And then we select the sample(s) from these subgroups:



Here, we first divided our population into subgroups based on different colours of red, yellow, green and blue. Then, from each colour, we selected an individual in the proportion of their numbers in the population. We use this type of sampling when we want representation from all the subgroups of the population.

Cluster Sampling

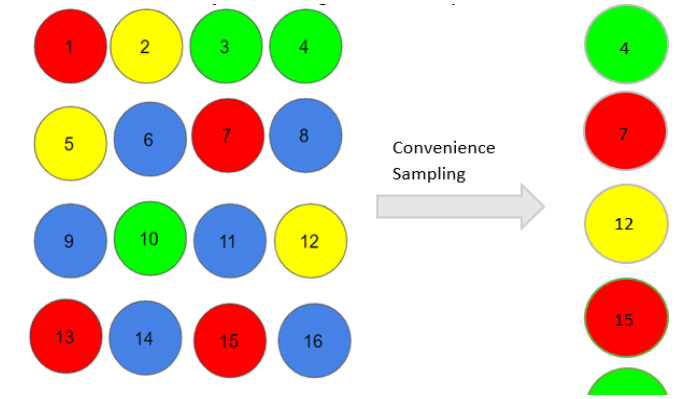
In a clustered sample, we use the subgroups of the population as the sampling unit rather than individuals. The population is divide clusters, and a whole cluster is randomly selected to be included in the study. We have divided our population into 5 clusters. Each cluster consists of 4 individuals and we have taken the 4th cluster in our sample. **This type of sampling is used when we focus on a specific region or area.**



Types of Non-Probability Sampling

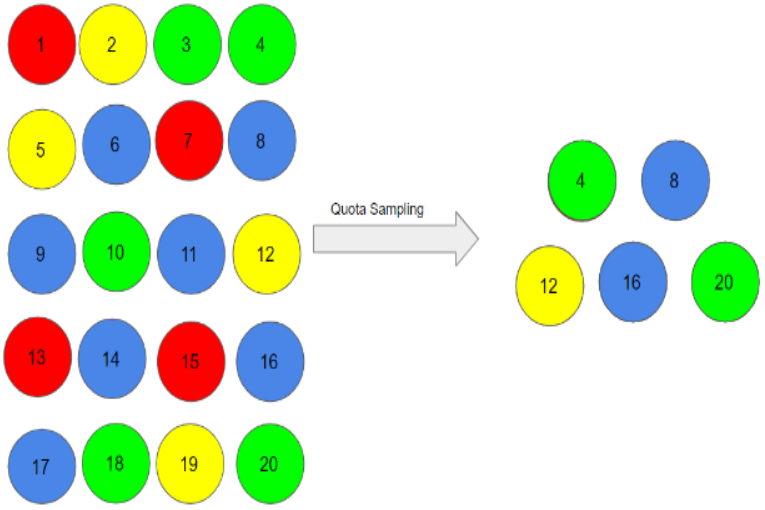
Convenience Sampling

This is perhaps the easiest method of sampling because individuals are selected based on their availability and willingness to take part. Let's say individuals numbered 4, 7, 12, 15 and 20 want to be part of our sample, and hence, we will include them in the sample. Convenience sampling is prone to significant bias, because the sample may not be the representation of the specific characteristics such as religion or, say the gender, of the population.



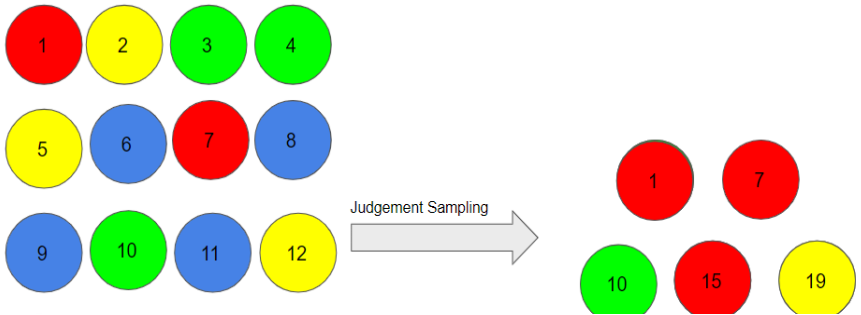
Quota Sampling

In this type of sampling, we choose items based on predetermined characteristics of the population. Consider that we have to select individuals having a number in multiples of four for our sample. Therefore, the individuals numbered 4, 8, 12, 16, and 20 are already reserved for our sample. In quota sampling, the chosen sample might not be the best representation of the characteristics of the population that weren't considered.



Judgment Sampling

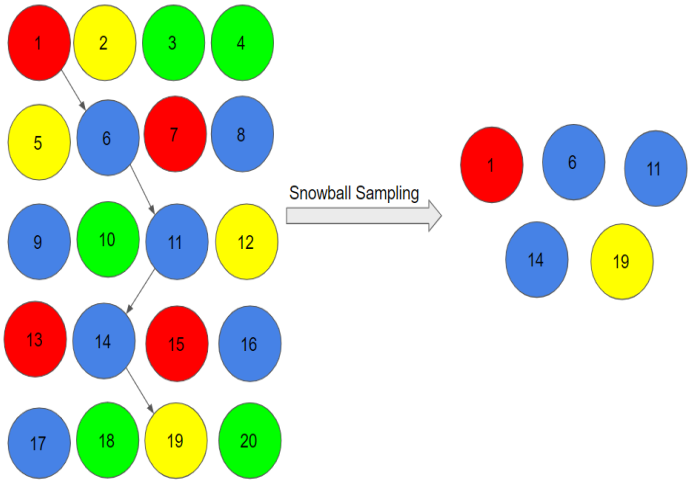
It is also known as ***selective sampling***. It depends on the judgment of the experts when choosing whom to ask to participate. Suppose, our experts believe that people numbered 1, 7, 10, 15, and 19 should be considered for our sample as they may help us to infer the population in a better way. As you can imagine, quota sampling is also prone to bias by the experts



and may not necessarily be representative.

Snowball Sampling

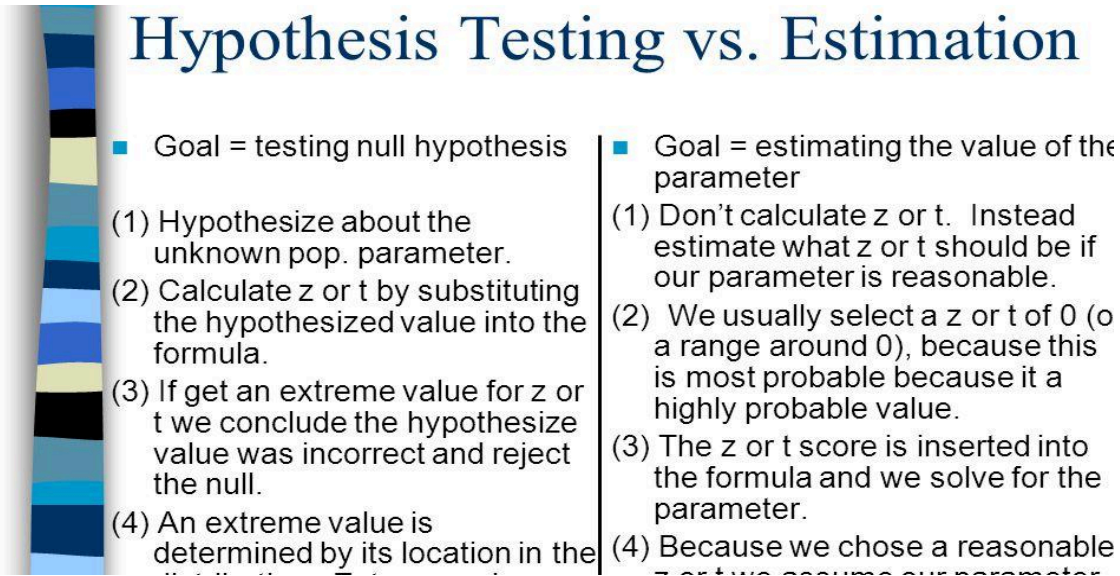
I quite like this sampling technique. **Existing people are asked to nominate further people known to them so that the sample increases in size like a rolling snowball.** This method of sampling is effective when a sampling frame is difficult to identify. Here, we had randomly chosen person 1 for our sample, and then he/she recommended person 6, and person 6 recommended person 11, and so on: **1->6->11->14->19** There is a significant risk of selection bias in snowball sampling, as the referenced individuals will share common traits with the person who recommends them.



Q17. What is the difference between parameter estimation and hypothesis testing?

Parameter estimation and hypothesis testing are two fundamental concepts in statistics, often used in statistical inference.

Parameter Estimation: Parameter estimation involves the process of estimating unknown parameters of a population based on sample data. The goal is to find the best guess or estimate for the values of one or more parameters that characterize the population.



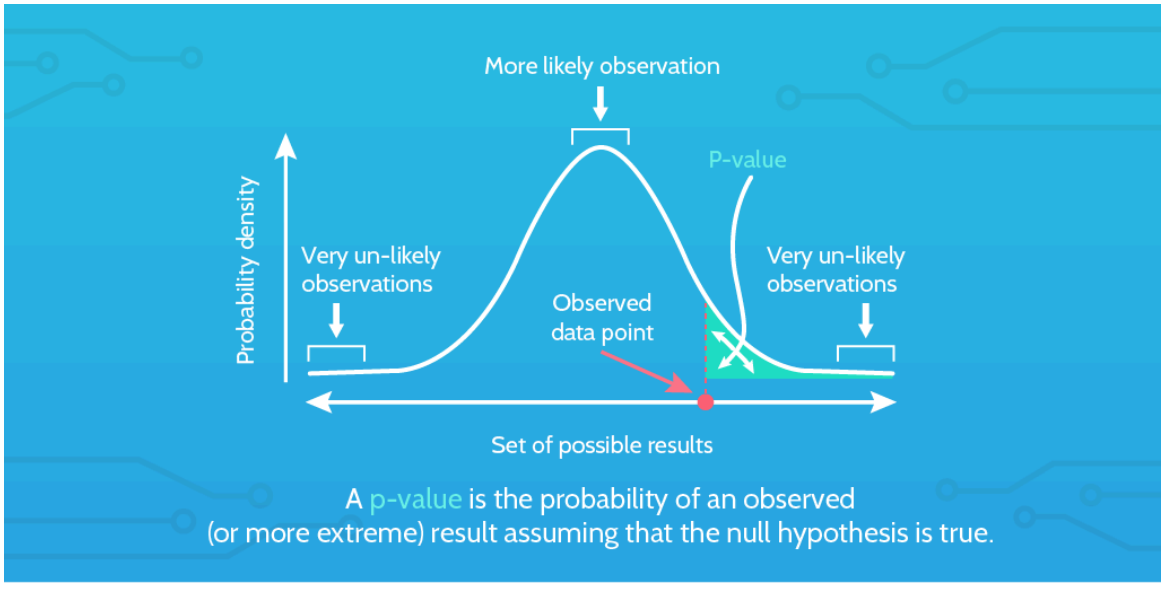
Hypothesis Testing: Hypothesis testing is a statistical method used to make inferences about a population parameter based on a sample of data.

In practice, these two concepts are often used together. For example, you might estimate a parameter and then conduct hypothesis testing to assess whether the estimated value is significantly different from a certain value or if there is a significant effect.

Q18. What is the p-value in hypothesis testing?

The P value is defined as the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed. The P stands for probability and measures how likely it is that any observed difference between groups is due to chance.

Being a probability, P can take any value between 0 and 1. Values close to 0



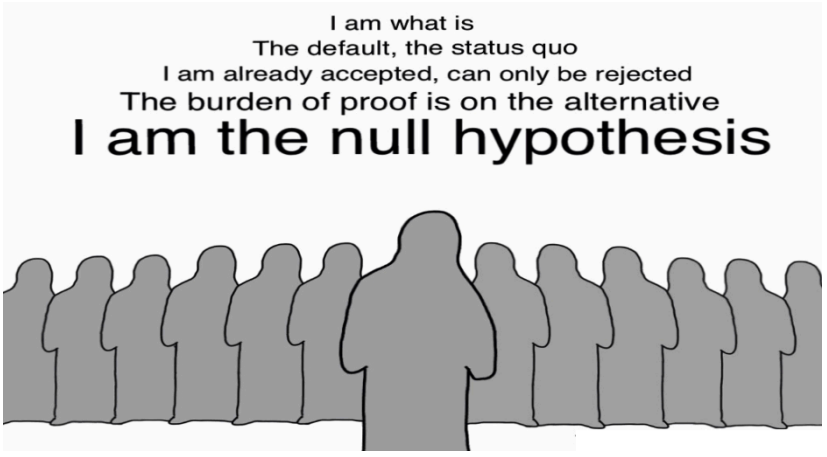
indicate that the observed difference is unlikely to be due to chance, whereas a P value close to 1 suggests no difference between the groups other than due to chance. Thus, it is common in medical journals to see adjectives such as “highly significant” or “very significant” after quoting the P value depending on how close to zero the value is.

Small p value(less than 0.05): Indicates strong evidence against the null hypothesis. You may reject the null hypothesis in favour of the alternative.

Large p value (larger than 0.05): Indicates weak evidence against the null hypothesis. You may fail to reject the null hypothesis.

How p-values work in hypothesis testing

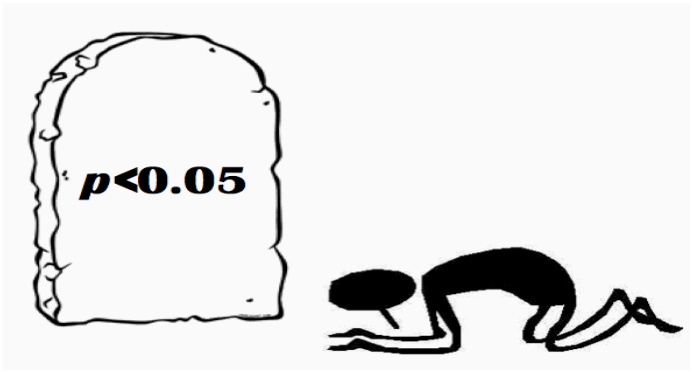
Null hypothesis (H_0): This is a statement of no effect or no difference. It represents the status quo or a default assumption.



Alternative hypothesis (H_a or H_1): this is a statement that contradicts the null hypothesis, suggesting the presence of an effect or a difference.

Test static: It is a numerical summary of the sample data, often standardized in a way that makes it comparable to a standard distribution (e.g., a z-score or t-statistic).

P value: The p-value is the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the sample data, assuming the null hypothesis is true. A low p-value suggests that the observed data is unlikely if the null hypothesis is true, leading to the rejection of the null hypothesis.



Decision rule: Based on a pre-defined significance level (usually denoted as α, commonly set at 0.05), you compare the p-value to this threshold. If the p-value is less than or equal to α, you reject the null hypothesis. If the p-value is greater than α, you fail to reject the null hypothesis.

Q19. What is confidence interval estimation?

Confidence interval estimation is a statistical technique used to estimate a range of values within which a population parameter is likely to lie with a certain level of confidence. It provides a measure of the uncertainty associated with a point estimate of a parameter based on sample data.

Here are the key components and concepts related to confidence interval estimation:

Point estimate: A point estimate is a single value that serves as the best guess for the population parameter based on the sample data. Common point estimates include the sample mean for estimating the population mean and the sample proportion for estimating a population proportion.

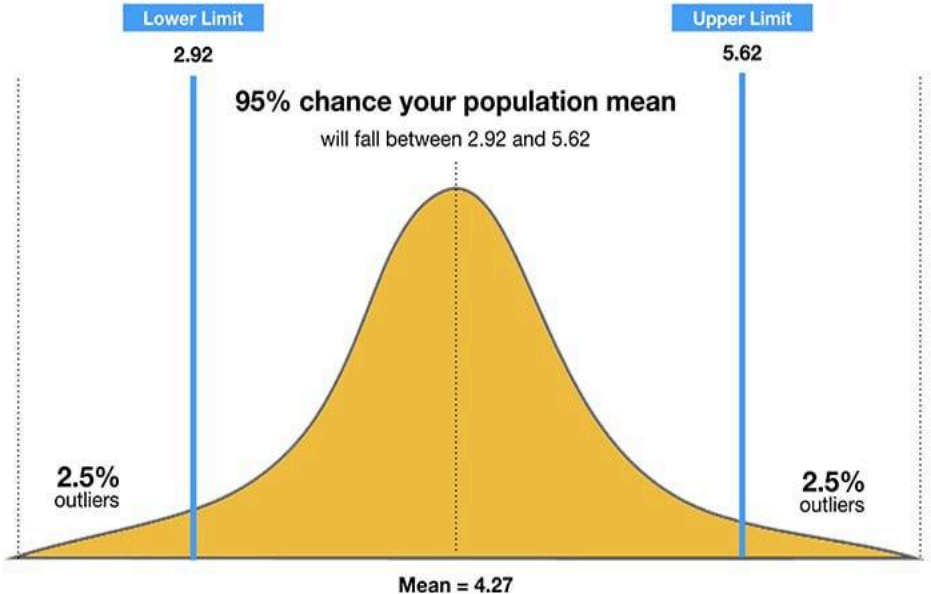
Margin of error: The margin of error is a measure of the variability or uncertainty associated with the point estimate. It is typically expressed as a range of values above and below the point estimate.

Confidence level: The confidence level is the probability that the confidence interval contains the true population parameter. It is often expressed as a percentage, such as 95% or 99%. A 95% confidence level, for example, implies that if we were to take many samples and construct a confidence interval for each, approximately 95% of those intervals would contain the true population parameter.

Formula for confidence interval: The general formula for a confidence interval is***: Confidence Interval=Point Estimate±Margin of Error.***

The margin of error is calculated based on the standard error of the point estimate and is influenced by the chosen confidence level.

Standard error: The standard error is a measure of the variability of the point estimate. It takes into account the sample size and the variability of the data.



Example:

Suppose you want to estimate the average height of a population. You collect a sample and calculate the sample mean. Using the standard error and the chosen confidence level (e.g., 95%), you construct a confidence interval. This interval represents a range of values within which you are reasonably confident the true population mean height resides.

Interpretation: If you construct a 95% confidence interval for a parameter, it means that if you were to repeat the sampling process many times, you would expect the true parameter to fall within the calculated interval about 95% of the time.

In conclusion, confidence interval estimation provides a way to quantify the uncertainty associated with a point estimate and helps convey the range of values within which the true population parameter is likely to exist. The choice of confidence level reflects the level of certainty desired by the researcher or analyst.

Q20. What are Type I and Type II errors in hypothesis testing?

In hypothesis testing, Type I and Type II errors are two types of mistakes that can occur when making decisions about a null hypothesis.

Type I error (False positive):

Definition: A Type I error occurs when you reject a null hypothesis that is actually true. In other words, it is the error of concluding that there is a significant effect or difference when there is none in the population.

Probability: The probability of committing a Type I error is denoted by the symbol α (alpha), and it is the chosen significance level (e.g., 0.05 or 5%). The lower the significance level, the lower the chance of making a Type I error, but it increases the risk of Type II error.

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β

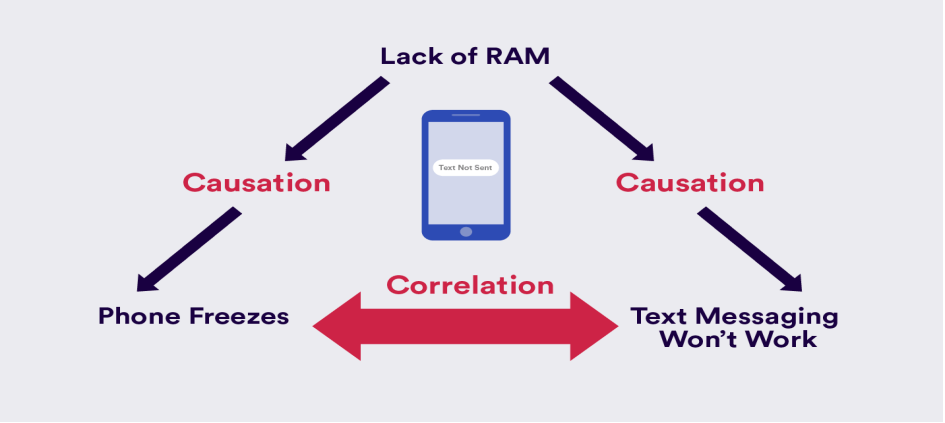
Type II error (False negative):

Definition: A Type II error occurs when you fail to reject a null hypothesis that is actually false. It is the error of not detecting a real effect or difference when one exists in the population.

Probability: The probability of committing a Type II error is denoted by the symbol β (beta). Power of the test, which is $1-1-\beta$, is the probability of correctly rejecting a false null hypothesis. The power of a test is influenced by factors such as sample size, effect size, and the chosen significance level.

Q21. What is the difference between correlation and causation?

My mother recently complained to me: “Whenever I try to text message, my phone freezes.” A quick look at her smartphone confirmed my suspicion: she had five game apps open at the same time plus Facebook and YouTube. The act of trying to send a text message wasn’t causing the freeze, the lack of RAM was. But she immediately connected it with the last action she was doing before the freeze. She was implying a causation where there was only a correlation.



Let’s understand what is correlation and causation:

Correlation: Correlation is a term in statistics that refers to the degree of **association between two random variables**. So the correlation between two data sets is the amount to which they resemble one another.

There are three types of correlations that we can identify:

- **Positive correlation** is when you observe A increasing and B increases as well. Or if A decreases, B correspondingly decreases. Example: the more purchases made in your app, the more time is

spent using your app. **Negative correlation** is when an increase in A leads to a decrease in B or vice versa.

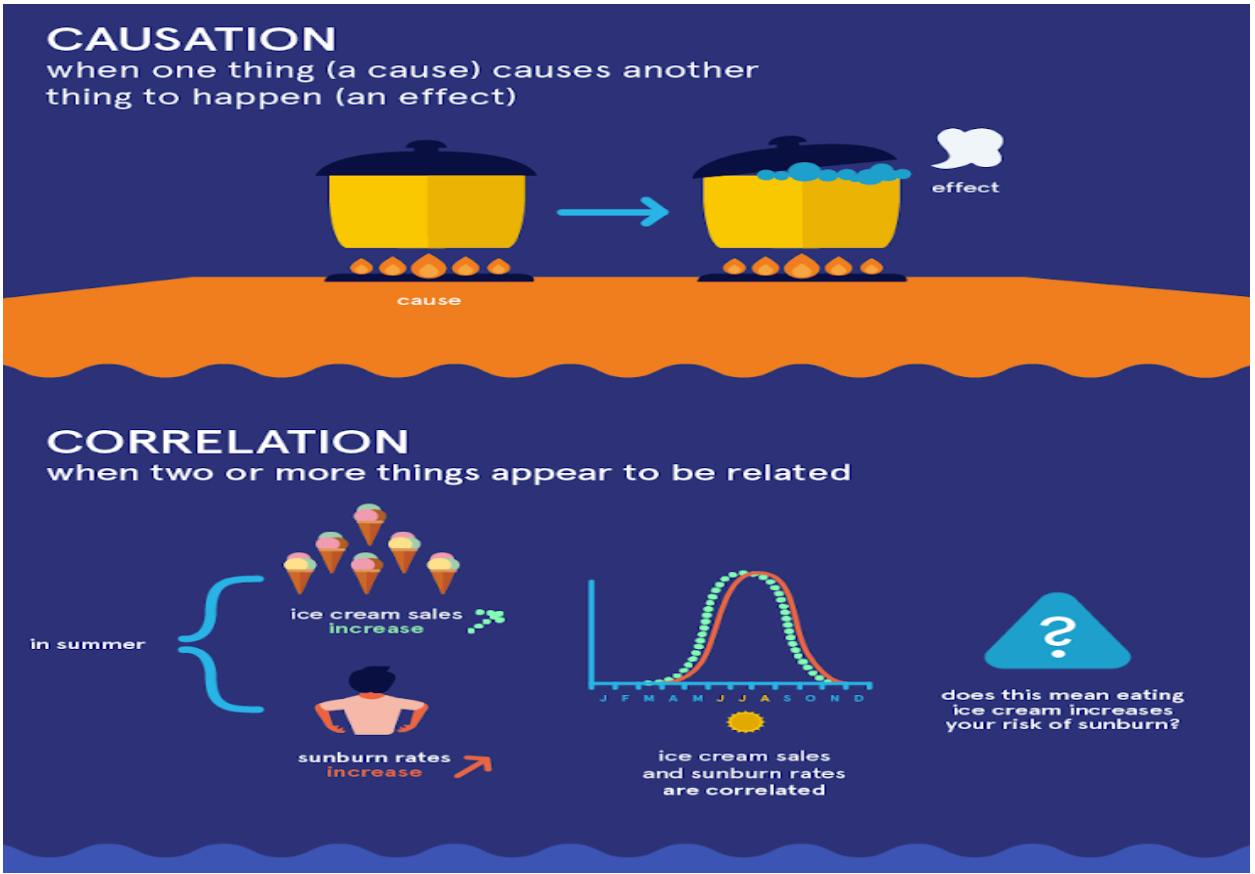
- **No correlation** is when two variables are completely unrelated and a change in A leads to no changes in B, or vice versa.

Causation: Causation is implying that **A and B have a cause-and-effect** relationship with one another. You're saying A causes B. Causation is also known as causality.

Firstly, causation means that two events appear at the same time or one after the other. And secondly, it means these two variables not only appear together, the existence of one causes the other to manifest.

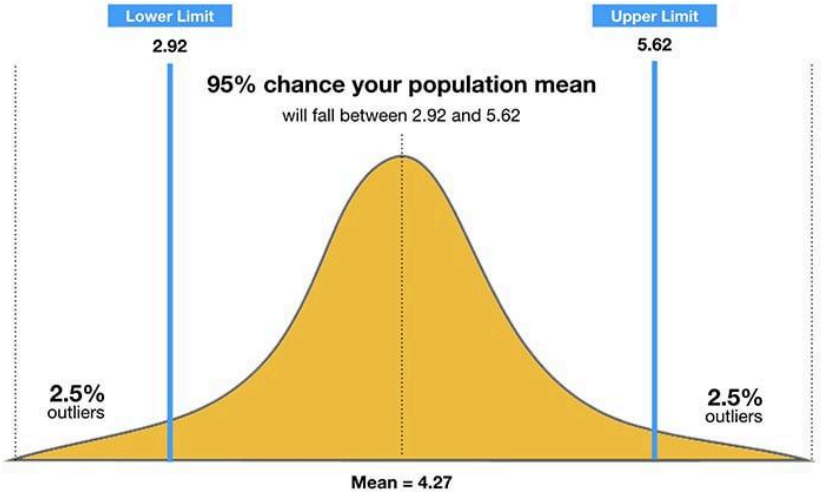
Difference between Correlation and Causation

• Warmer weather caused more sales of icecreams	• The weather gets warmer with increase in sales of icecreams
• Too many steps in purchasing on an app leads to cart abandonment and also uninstalling the app	• Cart abandonment means more people Uninstalling the app.
• Smokers like to drink coffee and smokers develop lung cancer	• Coffee drinkers are at high risk of lung cancer.
• Stress cuases Depression	• Stress cuases rumination which causes depression.



Q22. How is a confidence interval defined in statistics?

A confidence interval (CI) is a statistical concept used to estimate the range within which a population parameter is likely to lie, based on a sample of data. It provides a range of values that is believed to contain the true value of the parameter with a certain level of confidence. The confidence interval is expressed as a range and is associated with a confidence level, typically expressed as a percentage. Confidence intervals show the degree of uncertainty or certainty in a sampling method. They are constructed using confidence levels of 95% or 99%.



The 95% confidence interval is the range that you can be 95% confident that the similarly constructed intervals will contain the parameter being estimated. The sample mean (centre of the CI) will vary from sample to sample because of natural sampling variability.

The formula to find Confidence Interval is:

$$\bar{x} \pm Z \frac{s}{\sqrt{n}}$$

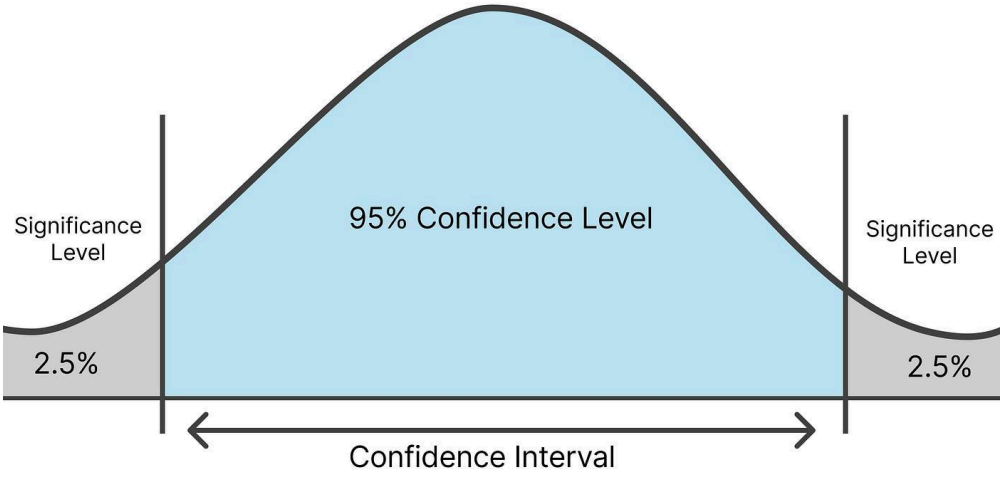
- X bar is the sample mean.
- Z is the number of standard deviations from the sample mean.
- S is the standard deviation in the sample.
- n is the size of the sample.

The value after the ± symbol is known as the margin of error.

Q23. What does the confidence level represent in a confidence interval?

The confidence level in a confidence interval represents the probability or likelihood that the interval will contain the true population parameter. It is a measure of the reliability of the interval estimation. Commonly used confidence levels include 90%, 95%, and 99%, with 95% being the most widely used.

When you construct a confidence interval, you are essentially saying, "I am X% confident that the true parameter lies within this interval." For example, if you construct a 95% confidence interval for the mean, it means that if you were to take many samples from the same population and calculate a confidence interval for each sample, you would expect approximately 95% of those intervals to contain the true population mean.



If you were to construct 100 different 95% confidence intervals from 100 different samples, you would expect around 95 of them to contain the true population parameter, and about 5 of them would not.

Confidence levels and intervals are important because they help statisticians understand the probability that a parameter is between values around the mean. These measurements can help represent degrees of certainty regarding surveys and study results. They help statisticians understand how likely it is that they can receive the same results each time they complete a study. The choice

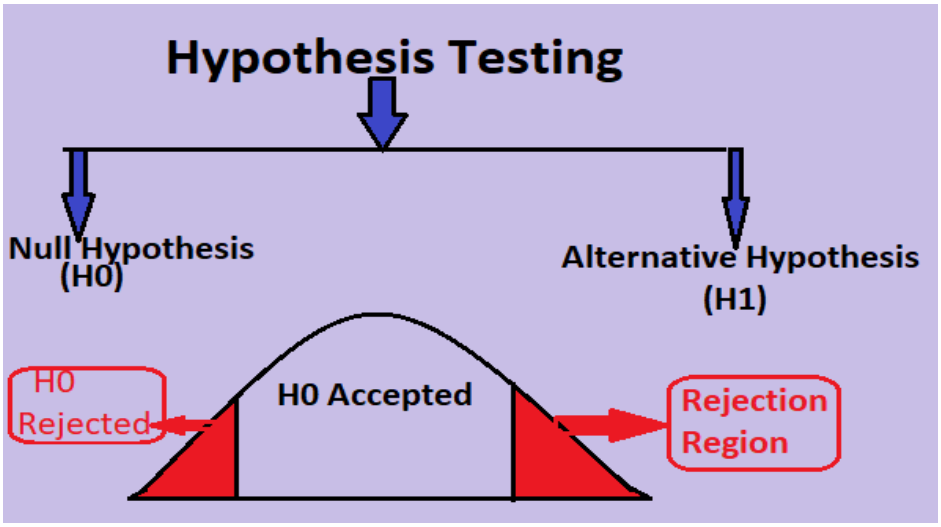
of confidence level depends on the level of certainty required for a particular application or decision-making process.

Q24. What is hypothesis testing in statistics?

Hypothesis testing is a statistical method used to make inferences about population parameters based on a sample of data. The process involves formulating a hypothesis, collecting and analysing data, and drawing conclusions about the population based on the results. It is used to estimate the relationship between two statistical variables.

Let's discuss few examples of statistical hypothesis from real-life –

- A teacher assumes that 60% of his college's students come from lower-middle-class families.
- A Doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.



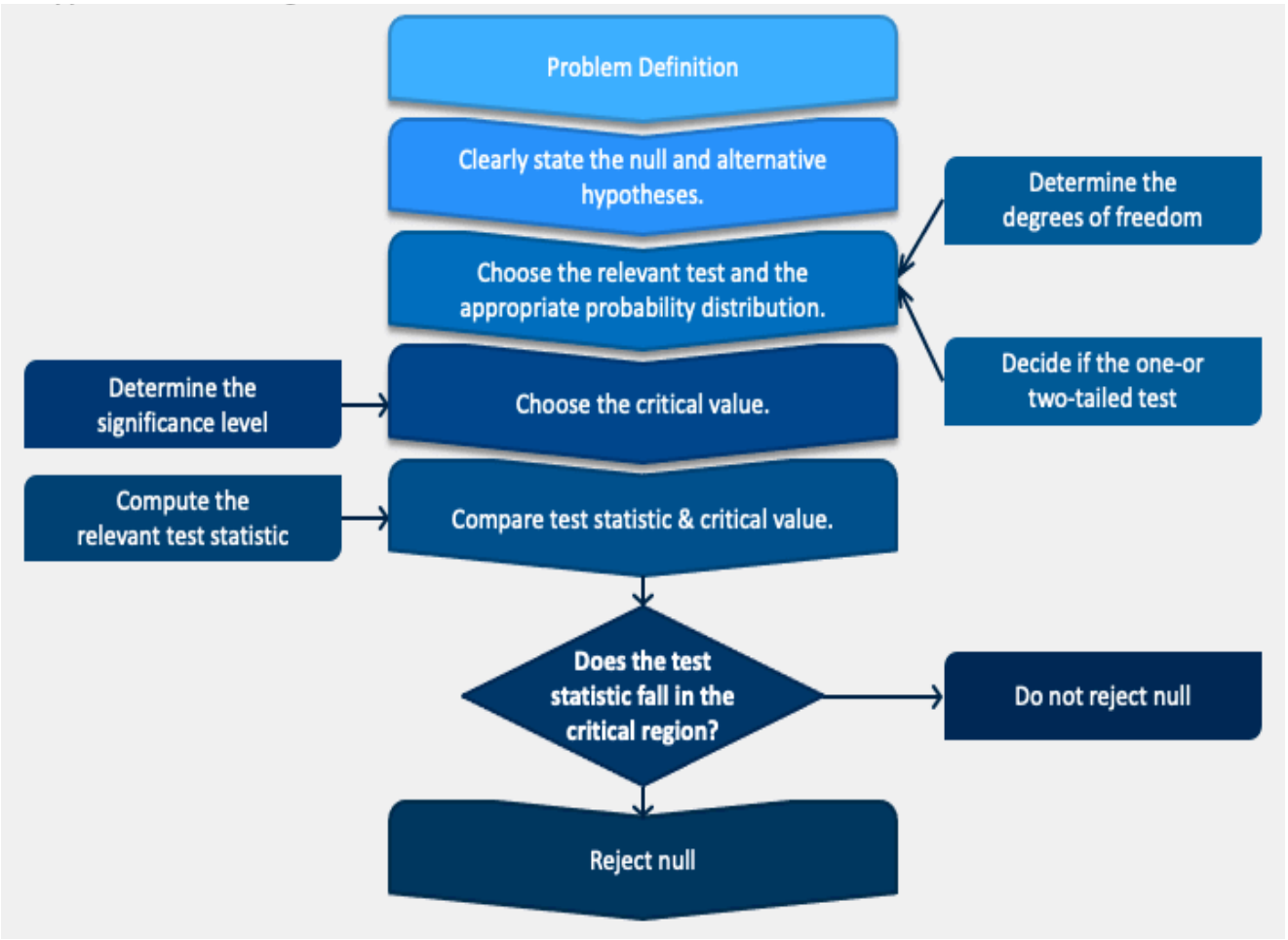
Null Hypothesis and Alternate Hypothesis

The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.
H0 is the symbol for it, and it is pronounced H-naught.

The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. H1 is the symbol for it.
A sanitizer manufacturer claims that its product kills 95 percent of germs on average.
To put this company's claim to the test, create a null and alternate hypothesis.

H0 (Null Hypothesis): Average = 95%.
Alternative Hypothesis (H1): The average is less than 95%.

Steps in hypothesis testing



Conclusion

If you reject the null hypothesis, you may conclude that there is enough evidence to support the alternative hypothesis.
If you fail to reject the null hypothesis, you may conclude that there is not enough evidence to reject the null hypothesis.

It's important to note that "failing to reject" the null hypothesis does not prove the null hypothesis to be true; it simply means that there is not enough evidence to reject it based on the available data. Hypothesis testing is a

fundamental tool in inferential statistics and is widely used in various fields to make decisions and draw conclusions about populations based on sample data.

Q25. What is the purpose of a null hypothesis in hypothesis testing?

The null hypothesis (*H*0) in hypothesis testing serves as a starting point or a baseline assumption. Its purpose is to represent a statement of no effect, no difference, or no change in the population parameter of interest. The null hypothesis essentially embodies the status quo or the idea that there is no real effect or relationship in the population.

Here are key purposes of the null hypothesis:

- ***Establishing a baseline:*** The null hypothesis provides a benchmark against which researchers can compare their findings. It assumes that any observed differences or effects in the sample are due to random variation or chance, rather than a genuine effect in the population.
- ***Formulating a testable statement:*** The null hypothesis is a testable statement that can be evaluated based on sample data. It allows researchers to set up a structured framework for hypothesis testing, where they can assess the likelihood of observing the results obtained if the null hypothesis were true.
- ***Defining the null distribution***: The null hypothesis helps define the null distribution, which represents the distribution of test statistics that would be expected if there were no real effect in the population. This distribution is crucial for determining the statistical significance of observed results.
- ***Facilitating statistical testing***: Hypothesis testing involves comparing observed data to what would be expected under the assumption that the null hypothesis is true. By specifying a null hypothesis, researchers can use statistical methods to assess whether the observed results are unlikely to occur by random chance alone.
- ***Setting the basis for inference****:* The null hypothesis is a foundation for making inferential decisions. When researchers perform hypothesis testing, they can either reject the null hypothesis in favour of the

alternative hypothesis or fail to reject the null hypothesis. This decision-making process guides conclusions about the population based on sample data.

Q26. What is the difference between a one-tailed and a two-tailed test?

A one-tailed test and a two-tailed test refer to different ways of setting up and analysing the results of a statistical hypothesis test. The distinction lies in the directionality of the test and the focus on specific regions of the probability distribution.

Parameters	One-Tailed	Two-Tailed
What are They?	A one-tailed test is a method of hypothesis testing that only looks for an effect in one direction based on a prior hypothesis.	A two-tailed test is a method of hypothesis testing that looks for an effect in both directions without a prior hypothesis.
Purpose	The purpose of a one-tailed test is to test for an effect in a specific direction based on a prior hypothesis (e.g. students who study more hours will have higher grades)	The purpose of a two-tailed test is to test for an effect in any direction without a prior hypothesis (e.g. there is a difference between the grades of male and female students)
Critical value	A one-tailed test typically has a smaller critical value than a two-tailed test.	A two-tailed test typically has a larger critical value than a one-tailed test.
Alpha level	The alpha level associated with a one-tailed test is typically larger than that associated with a two-tailed test.	The alpha level associated with a two-tailed test is typically smaller than that associated with a one-tailed test.
Interpretation of Results	The interpretation of results from a one-tailed test tends to be more straightforward as it only tests for effects in one direction.	Two-tailed tests can detect effects in multiple directions, making interpretation more complicated.
Sample Size	The sample size required for a one-tailed test is typically larger than the sample size required for a two-tailed test.	The sample size required for a two-tailed test is typically smaller than one-tailed test.
Hypotheses	With a one-tailed hypothesis we would ask whether "group A scores higher than group B"	With a two-tailed hypothesis we would ask whether "there is a difference between the scores of group A and group B".

Q27. What is experiment design, and why is it important?

Experimental design refers to the process of planning and organizing an experiment to obtain valid and reliable results. It involves making decisions about the conditions under which the experiment will be conducted, the variables to be manipulated and measured, and the methods to be used. A well-designed experiment allows researchers to draw meaningful conclusions, establish cause-and-effect relationships, and generalize findings to a larger population.

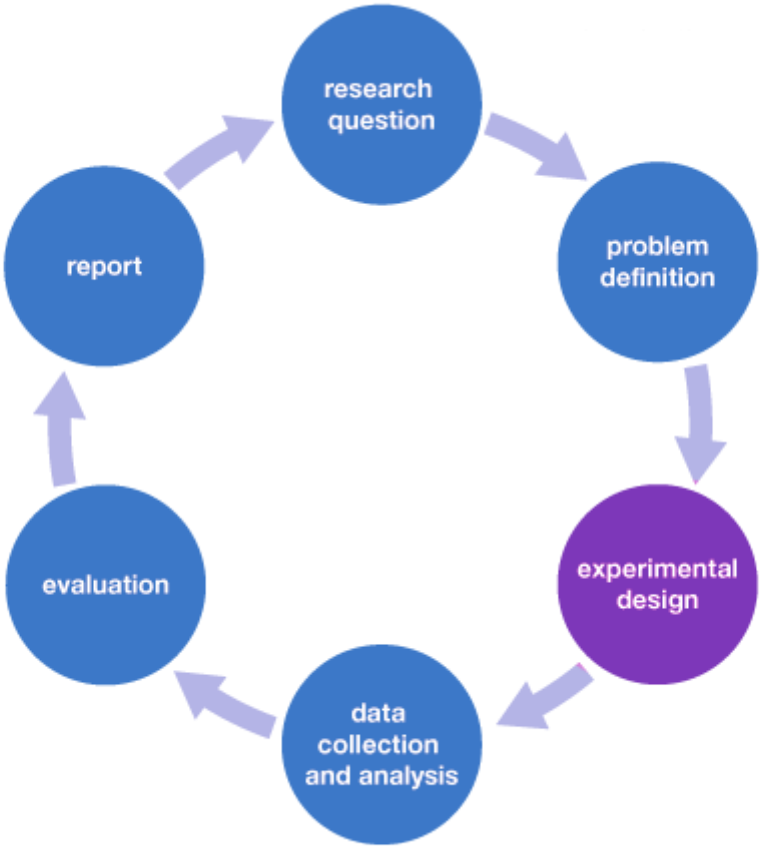
Importance of experiment design:

- ***Causation***: Well-designed experiments provide a basis for establishing cause-and-effect relationships between variables, allowing researchers to make meaningful conclusions about the impact of the independent variable on the dependent variable.
- ***Validity***: Proper experimental design enhances the validity of study results by minimizing biases and controlling for extraneous variables. This ensures that the findings accurately reflect the effects of the manipulated variable.
- ***Reliability:*** Reproducibility and consistency in results are essential for the reliability of a study. A carefully designed experiment helps achieve reliable outcomes that can be trusted and replicated by other researchers.
- ***Efficiency:*** A well-designed experiment maximizes the efficiency of data collection and analysis, saving time and resources. This allows researchers to answer their research questions effectively.
- ***Generalizability:*** Experimental design considerations, such as randomization and appropriate sampling, contribute to the external validity of the study, allowing researchers to generalize their findings to broader populations.

In summary, experimental design is critical for conducting scientifically rigorous research. It ensures that experiments are well-controlled, valid, and reliable, ultimately contributing to the advancement of knowledge in various fields.

Q28. What are the key elements to consider when designing an experiment?

Designing a successful experiment requires careful consideration of various key elements to ensure that the study is well-controlled, valid, and reliable. ***Here are the key elements to consider when designing an experiment:***



Research question and hypothesis: Clearly define the research question or problem that the experiment aims to address.
Formulate a testable hypothesis that predicts the relationship between the independent and dependent variables.

Independent and dependent variables: Identify the independent variable (the variable manipulated by the researcher) and the dependent variable (the variable measured to assess the effect of the independent variable). Operationalize these variables by specifying how they will be measured or manipulated.

Control group: Include a control group that does not receive the experimental treatment. This provides a baseline for comparison and helps isolate the effects of the independent variable.

Randomization: Randomly assign participants to different experimental conditions. Randomization helps control for individual differences and ensures that any observed effects are likely due to the manipulation.

Counterbalancing: If there are multiple conditions, use counterbalancing to distribute the order of presentation of conditions across participants. This helps control for order effects, such as learning or fatigue.

Binding: Consider blinding techniques, such as single-blind or double-blind procedures, to minimize biases. Blinding helps prevent expectations or preferences from influencing results, both in participants and researchers.

Sample size: Determine an appropriate sample size to achieve sufficient statistical power. A larger sample size increases the likelihood of detecting true effects and enhances the generalizability of findings.

Validity and reliability: Maximize internal validity by designing the experiment to measure what it intends to measure. Ensure that the study has external validity by considering how the findings can be generalized to other populations or settings. Strive for reliability by using consistent methods and measures.

Ethical considerations: Adhere to ethical guidelines in participant recruitment, informed consent, data collection, and reporting of results. Ensure that the experiment does not cause harm to participants and is conducted with integrity.

Pilot testing: Conduct a pilot study to test the feasibility of the experimental design. This helps identify and address potential issues before conducting the main experiment.

Feedback and iteration: Seek feedback from colleagues or experts in the field to improve the experimental design. Be open to iteration and refinement based on feedback and preliminary results.

By carefully addressing these key elements, researchers can enhance the quality of their experimental designs and increase the likelihood of obtaining valid and reliable results.

Q29. How can sample size determination affect experiment design?

Sample size determination plays a crucial role in experiment design, influencing various aspects of the study. The size of the sample (i.e., the number of participants or units included in the study) has implications for the statistical power of the experiment, the precision of the results, and the generalizability of findings.

Here are some ways in which sample size determination can affect experiment design:

Statistical power: Statistical power is the probability that a study will correctly reject a false null hypothesis. Increasing the sample size generally enhances statistical power. A larger sample size provides greater sensitivity to detect true effects if they exist.

Experimenters often conduct power analyses to determine the minimum sample size needed to achieve a desired level of power. This analysis considers factors such as effect size, significance level, and variability in the data.

Precision and confidence interval: A larger sample size leads to narrower confidence intervals around the estimated effects. Narrower confidence intervals indicate greater precision in estimating the population parameters. Precision is important for drawing accurate and reliable conclusions. A more precise estimate allows researchers to have greater confidence in the range within which the true population parameter is likely to fall.

Effect size detection: The ability to detect small or subtle effects is influenced by the sample size. Larger sample sizes increase the likelihood of detecting smaller effect sizes, which may be of practical or theoretical importance. Researchers should consider the minimum effect size they want to detect and use this information in determining an appropriate sample size.

Type I and Type II errors: Sample size affects the balance between Type I and Type II errors. Increasing the sample size reduces the risk of Type II errors (false negatives) but may increase the risk of Type I errors (false positives) if not appropriately adjusted.

Resource allocation: The practical feasibility of the study is influenced by the available resources, including time, funding, and personnel. Larger sample sizes may require more resources, both in terms of data collection and analysis.

Generalizability: The size and diversity of the sample impact the generalizability of study findings to the broader population. A more representative sample enhances the external validity of the study. Researchers should consider whether the sample adequately represents the population of interest and whether the findings can be generalized beyond the study sample.

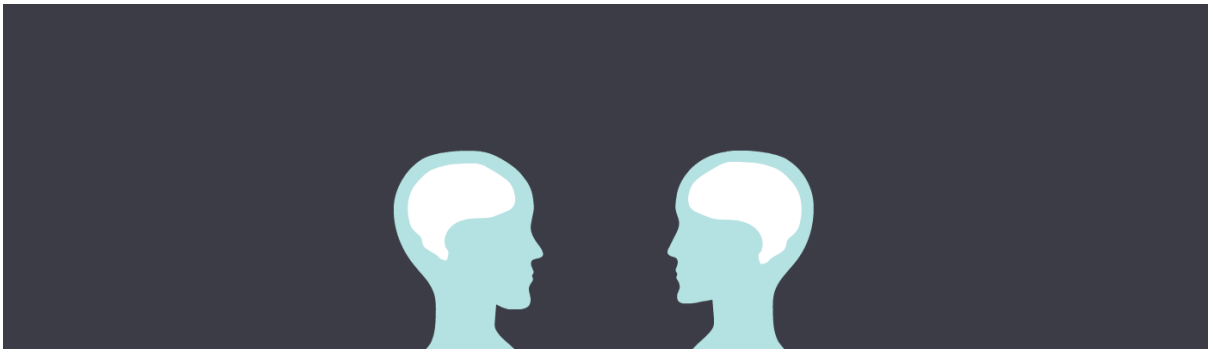
Ethical considerations: Ethical considerations, such as the potential burden on participants, informed consent, and the risk-benefit ratio, are relevant when determining sample size. Researchers must balance the need for a sufficiently large sample with ethical considerations.

In conclusion, sample size determination is a critical aspect of experiment design, influencing statistical power, precision, effect size detection, error rates, resource allocation, generalizability, and ethical considerations. Researchers should carefully consider these factors to design experiments that are both scientifically rigorous and ethically sound.

Q30. What are some strategies to mitigate potential sources of bias in experiment design?

We've all experienced some form of bias in one way or the other. You may have seen it happen to others, experienced it yourself or even participated in it. Bias here means favouring something over another even when the thing being favoured does not deserve to be. Aside from our everyday lives, bias also occurs during experiments and research.

Bias in experiments refers to a known or unknown influence in the experimental process, data or results.



Sources of bias in experiments.

1. The method of data collection and the source of the data can lead to bias in experiments. To learn about the methods of data collection, see the article on Methods of Data Collection.
2. Not considering all possible outcomes can lead to bias. Even though, it is not really possible to consider all outcomes, scientist should make an effort to perform more experiments to control any new source of bias found.
3. Unknown changes in the experimental environment can lead to bias.
4. False behaviour and response from the participants can lead to bias.

Strategies to mitigate bias in experimental design:

1. Ensure that the participants in your experiment represent all categories that are likely to benefit from the experiment.
2. Ensure that no important findings from your experiments are left out.
3. Consider all possible outcomes while conducting your experiment.
4. Make sure your methods and procedures are clean and correct.
5. Seek the opinions of other scientists and allow them review you experiment. They may be able to identify things you have missed.
6. Collect data from multiple sources.
7. Allow participants to review the conclusion of your experiment so they can confirm that the conclusion accurately represents what they portrayed.
8. The hypothesis of an experiment should be hidden from the participants so they don't act in favour or maybe against it.
9. Implement single-blind or double-blind procedures to minimize bias due to expectations. In a single-blind study, participants are unaware of the treatment conditions, while in a double-blind study, both participants and experimenters are unaware. Blinding helps prevent conscious or unconscious biases in data collection and analysis.

By incorporating these strategies into the experimental design process, researchers can enhance the validity and reliability of their studies, reduce biases, and contribute to the overall rigor of scientific research.

Conclusion:

1. The results and conclusion of the experiment will be reliable and dependable.
2. There will be better chances of the experiment helping as much people as it should.
3. Important information and findings will not be hidden or left out.
4. The conclusion of the experiment will not be influenced by any specific opinion.
5. The scientist will be open minded and consider all possibilities while conducting the experiment.
6. The data collected will be more accurate.
7. Detailed and complete articles and journals for the experiment will be published.

Q31. What is the geometric interpretation of the dot product?

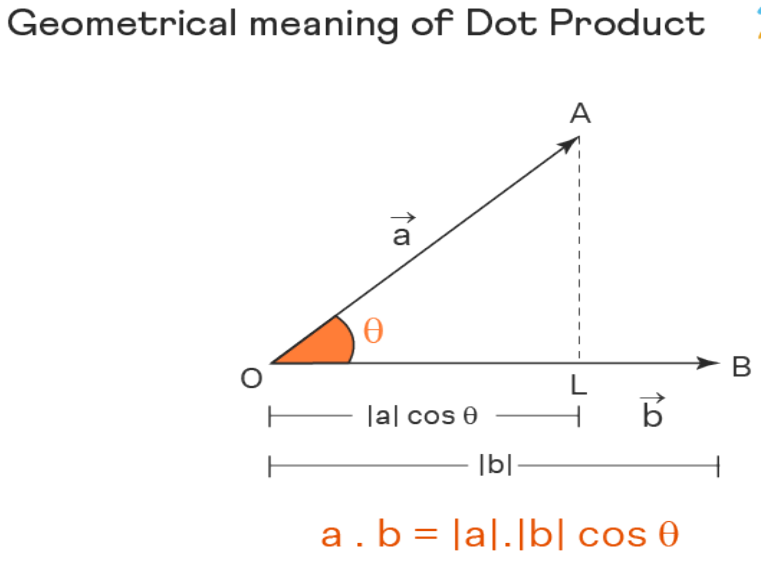
The dot product of two vectors is a mathematical operation that takes two equal-length sequences of numbers (vectors) and returns a single number. Geometrically, the dot product has a significant interpretation related to the angle between two vectors.

Given two vectors A and B, the dot product (also known as the scalar product) is calculated as follows:

$$A \cdot B = |A| \cdot |B| \cdot \cos(\theta)$$

Where,

- $A \cdot B$ is the dot product.
- $|A|$ and $|B|$ are the magnitudes (lengths) of vectors A and B, respectively.
- ϑ is the angle between vectors A and B.



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

- ***Parallel vectors ($\vartheta = 0^\circ$)*** When the vectors are parallel, the dot product is maximized, and the cosine of ϑ is 1. This means that $A \cdot B = |A| \cdot |B|$, and the vectors are pointing in the same direction.

- ***Perpendicular vectors ($\vartheta = 90^\circ$)*** 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 ($\theta = 180^\circ$)*** When the vectors are pointing in opposite directions, the dot product is minimized, and the cosine of ϑ is -1. This results in $A \cdot B = -|A| \cdot |B|$.

Geometrical Interpretation of Dot Product of Two Vectors

Suppose there are two vectors, P and Q.

$P = a_1i + a_2j + a_3k$, and $Q = b_1i + b_2j + b_3k$
Then, P.Q will define the scalar product.

$P.Q = (a_1i + a_2j + a_3k).(b_1i + b_2j + b_3k)$
$P.Q = \{(a_1b_1)*(i.i) + (a_1b_2)*(i.j) + (a_1b_3)*(i.k)\} + \{(a_2b_1)*(j.i) + (a_2b_2)*(j.j) + (a_2b_3)*(j.k)\} + \{(a_3b_1)*(k.i) + (a_3b_2)*(k.j) + (a_3b_3)*(k.k)\}$
$P.Q = a_1b_1 + a_2b_2 + a_3b_3$
So, 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 a vector a on vector b. It is like the shadow of vector a falling onto vector b.

- The projection of a vector a on the vector b is calculated by,

  $(a.b)/|b|$

- The projection of a vector b on the vector a is calculated by,

  $(a.b)/|a|$

- The angle between any two vectors formed by their intersection at one point is calculated by

  $Cos\vartheta = (a.b)/|a|*|b|$

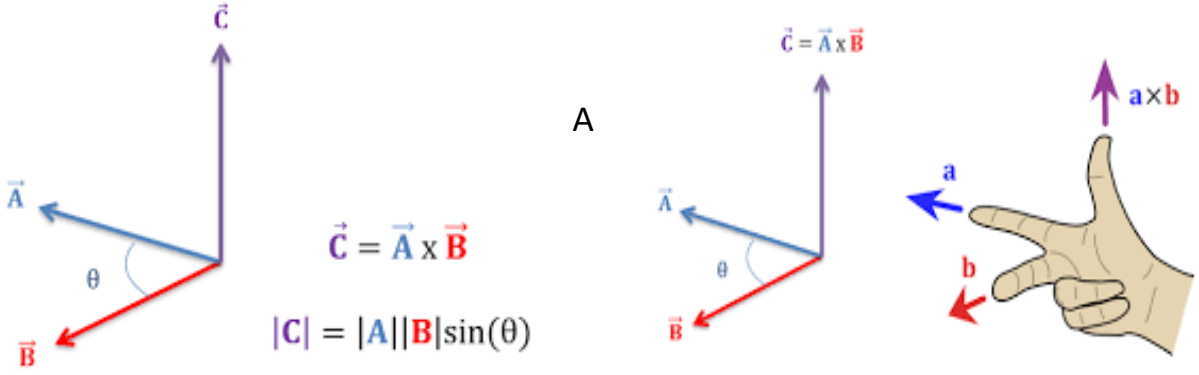
In summary, the dot product provides a measure of the similarity in direction of two vectors. It is positive for similar directions, zero for orthogonal vectors, and negative for vectors pointing in opposite directions. This geometric interpretation is useful in various

mathematical and physical applications, including physics, computer graphics, and linear algebra.

Q32. 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.***

The cross product is the other common name used to represent a vector multiplication. When two distinct or identical vectors have performed a multiplication, the result will also be a vector of different magnitude and net direction.



cross product is done if we consider more
about the direction part of the two vectors rather than the magnitudes of the two vectors. When we do the vector product of two vectors with different directions and magnitudes, it is evident that the direction of the resultant vector is found in the third dimension that is mutually perpendicular to the plane of the raw vectors.

Given two vectors **A**=(*A*1,*A*2,*A*3) and **B**=(*B*1,*B*2,*B*3), the cross product **A**×**B** is calculated as follows:

$$\mathbf{A}\times\mathbf{B}=(A2B3-A3B2, A3B1-A1B3, A1B2-A2B1)$$

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 = $(a_1i + a_2j + a_3k) \times (b_1i + b_2j + b_3k)$
P x Q = $(a_2b_3 - b_2a_3)i - (a_1b_3 - b_1a_3)j + (a_1b_2 - b_1a_2)k$
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. It is like the shadow of vector a falling onto vector b.

Key aspects:

- Direction: The resulting vector, $\mathbf{A}\times\mathbf{B}$, is perpendicular to the plane formed by vectors $\mathbf{A}$ and $\mathbf{B}$. The direction is determined by the right-hand rule, where you align your index finger with $\mathbf{A}$, your middle finger with $\mathbf{B}$, and your thumb points in the direction of $\mathbf{A}\times\mathbf{B}$.
- Magnitude: The magnitude of the cross product is equal to the area of the parallelogram formed by vectors $\mathbf{A}$ and $\mathbf{B}$. It can be calculated using the formula: $|\mathbf{A}\times\mathbf{B}|=|A|\cdot|B|\cdot\sin(\theta)$

Where, $|A|$ and $|B|$ are the magnitudes of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively, and θ is the angle between $\mathbf{A}$ and $\mathbf{B}$.

Conclusion: The geometrical interpretation of vectors is essential for the detailed understanding of why we consider projections while solving the problems and the impact it can lay to imagine quickly and come up with a better and quick solution-oriented method while solving vectors-related problems. Since vectors are the quantities that define both magnitude and the direction the object is heading. And vectors can exist not only in a single plane or two-dimensional plane of the paper, but they can also be represented in a three-dimensional plane which is why a vector has components associated with

it in all three different directions. These directions are the primary reason for the need to get a gold hold of the geometrical interpretation of vectors.

Q33. How are optimization algorithms with calculus used in training deep learning models?

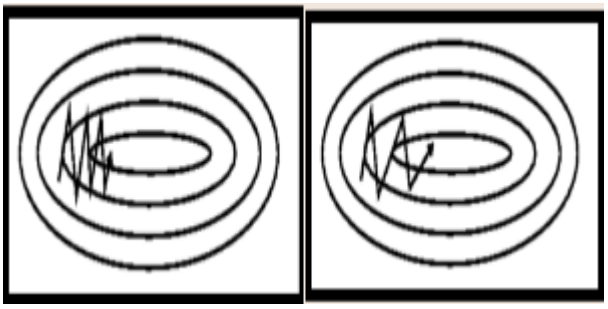
Optimization algorithms with calculus play a crucial role in training deep learning models. The goal of training a deep neural network is to find the set of model parameters that minimizes a certain objective function, often referred to as the loss function or cost function. Calculus-based optimization algorithms are employed to iteratively update the model parameters in the direction that reduces the value of this objective function.

Here's an overview of the key concepts involved:

- ***Loss function***: The loss function quantifies the difference between the predicted output of the model and the actual target values. The goal is to minimize this function during the training process.

- ***Gradient descent:*** Gradient descent is a first-order optimization algorithm that uses the gradient of the loss function with respect to the model parameters to iteratively update the parameters. The gradient points in the direction of the steepest increase of the function. The update rule for gradient descent is given by:
  new parameter=old parameter−learning rate×gradient
  The learning rate determines the step size in each iteration.

- ***Stochastic Gradient descent (SGD):*** In practice, the training dataset is often large, and computing the gradient using the entire dataset can be computationally expensive. SGD is a variant of gradient descent that randomly selects a subset (mini-batch) of the training data for computing the gradient and updating the parameters. This introduces stochasticity and can lead to faster convergence. ***In the above image, the left part shows the convergence graph of the stochastic gradient descent algorithm.*** At the



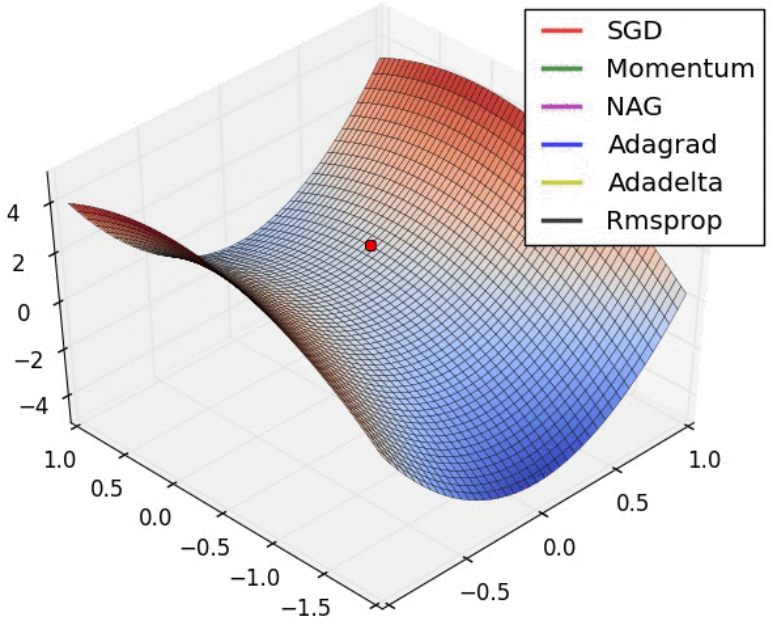
same time, the right side shows SGD with momentum. From the image, you can compare the path chosen by both algorithms and realize that using momentum helps reach convergence in less time.

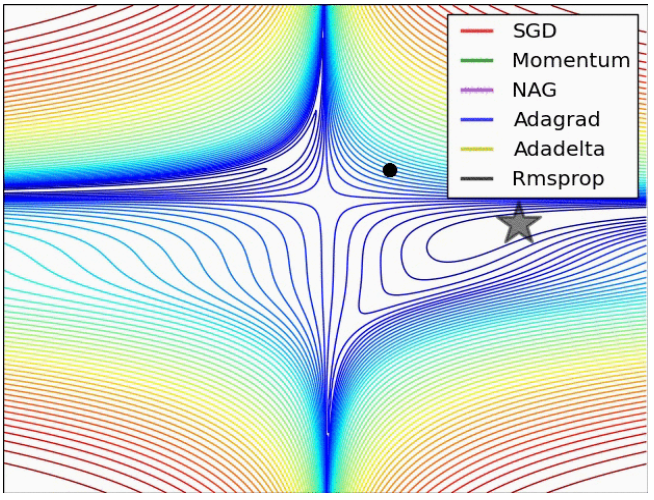
- ***Backpropogation***: Backpropogation is a technique used to efficiently compute the gradient of the loss function with respect to the model parameters. It leverages the chain rule of calculus to propagate the error backward through the network layers.

- ***Optimization algorithms:*** Various advanced optimization algorithms build upon the basic principles of gradient descent.

Examples include:

- **Adam:** (Adaptive Moment Estimation optimizer) A popular algorithm that combines ideas from momentum and RMSprop. It adapts the learning rates for each parameter individually.
- **Adagrad, Adadelta, RMSprop:** Other adaptive learning rate algorithms that aim to address the challenges of choosing a global learning rate.





The above visualizations create a better picture in mind and help in comparing the results of various optimization algorithms.

- ***Regularization:*** Calculus-based optimization allows the incorporation of regularization terms in the objective function, such as L1 or L2 regularization, to prevent over fitting.

The overall training process involves repeatedly updating the model parameters using the optimization algorithm until the loss function converges

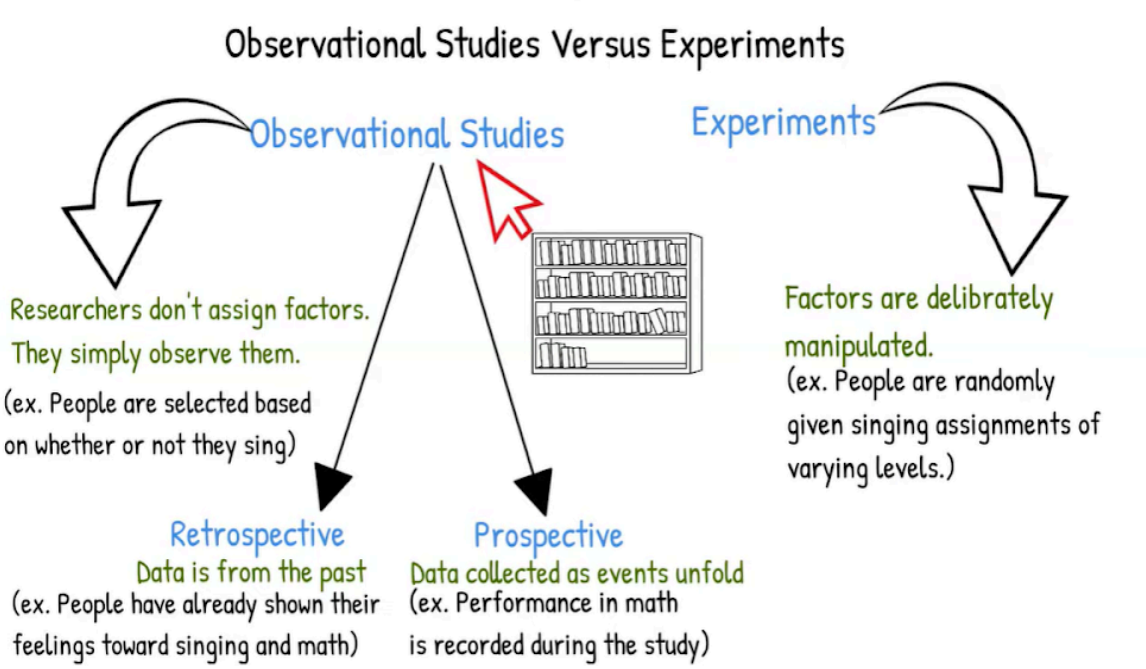
to a minimum or a satisfactory point. The use of calculus in computing gradients and determining the direction of parameter updates is fundamental to the success of training deep learning models.

Q34. What are observational and experimental data in statistics?

Observational and experimental data are two types of data collection approaches in statistics, each with its own characteristics and implications for drawing conclusions about relationships and causation.

Observational data: Observational data is collected by observing and recording the characteristics or behaviours of subjects or phenomena without intervening or manipulating any variables.

Experimental data: Experimental data is collected through experiments where researchers intentionally manipulate one or more variables to observe the effect on another variable.

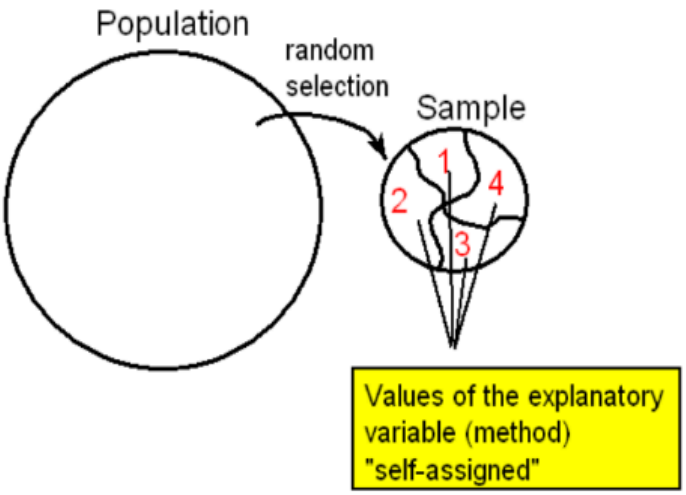


In an observational study, values of the explanatory variable occur naturally. In this case, this means that the participants themselves choose a method of trying to quit smoking. In an experiment, researchers assign the values of the explanatory variable. In other words, they tell people what method to use. In both cases, the goal is to gather information and draw

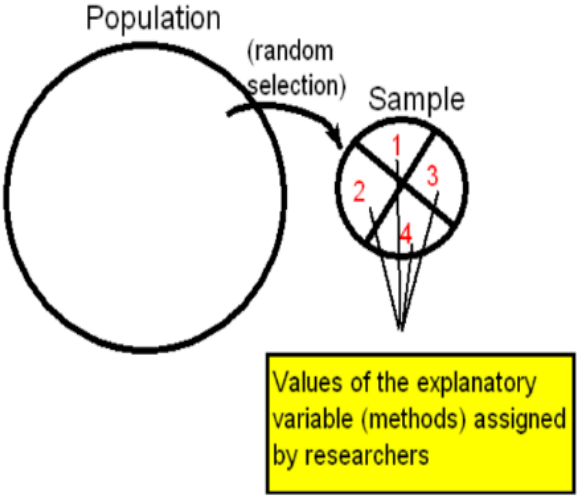
conclusions, but the methodology differs. Observational studies are more common in situations where experimentation is impractical or unethical, while experiments provide a stronger basis for establishing causation. Choosing the appropriate approach depends on the research question, ethical considerations, and practical constraints.

The following figures illustrate the two study designs:

Observational study



Experimental study



Examples:

Observational data: A researcher observes and records the eating habits of individuals in a cafeteria without intervening. The goal might be to understand the relationship between certain dietary choices and health outcomes.

Experimental data: A pharmaceutical company conducts a clinical trial where participants are randomly assigned to either a new drug (treatment group) or a placebo (control group). The goal is to assess the effectiveness of the drug in treating a specific condition.

Limitations:

Observational data: Because researchers do not control variables, establishing causal relationships is challenging. Confounding variables (factors that may affect the observed relationship) can be a concern.

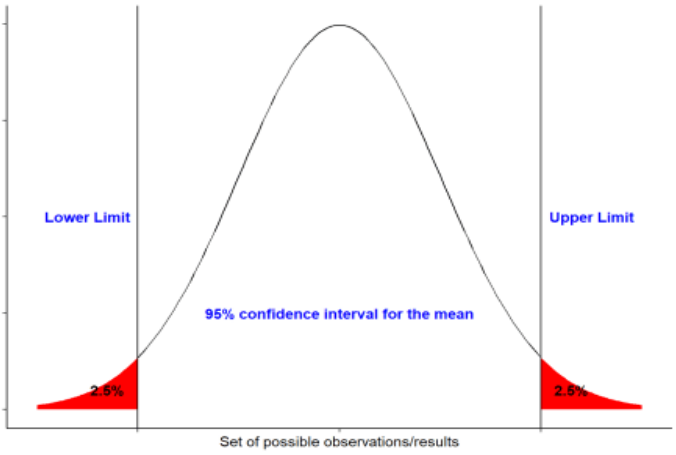
Experimental data: May be artificial or less reflective of real-world conditions. Ethical considerations may limit the types of experiments that can be conducted.

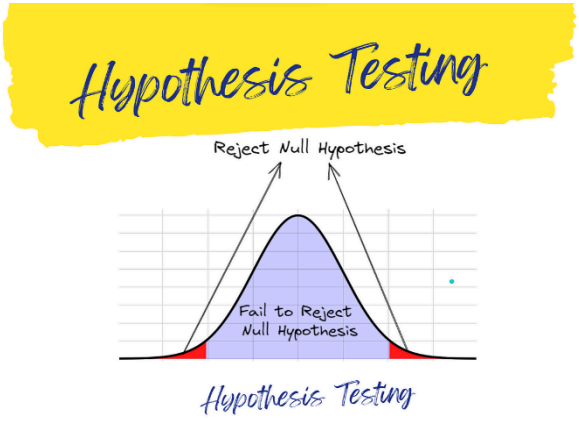
Q35. How are confidence tests and hypothesis tests similar? How are they different?

Confidence tests (intervals) and hypothesis tests are related concepts in statistics, both aimed at making inferences about population parameters based on sample data. While they share similarities, they serve different purposes and provide distinct types of information.

Similarities between Confidence tests and hypothesis tests:

- ***Inference from samples***: Both confidence tests and hypothesis tests involve making inferences about population parameters based on information obtained from a sample.
- ***Probability:*** They both involve the concept of probability. In hypothesis testing, p-values are used to assess the evidence against a null hypothesis, and in confidence intervals, the level of confidence is a measure of the probability that the interval contains the true population parameter.
- ***Sample data:*** Both approaches rely on sample data to draw conclusions about population parameters. The idea is to use the information from the sample to make inferences about the larger population.





While both confidence intervals and hypothesis tests involve statistical inference, they have different focuses and provide complementary information. Hypothesis tests assess the evidence against a specific hypothesis, while confidence intervals provide a range of values for a population parameter. Often, researchers use both approaches to gain a more comprehensive understanding of their data.

Difference between

	Hypothesis tests	***Confidence test***
Purpose	To assess the evidence against a specific null hypothesis. It helps determine whether there is enough evidence to reject the null hypothesis in favour of an alternative hypothesis.	To provide a range of plausible values for an unknown population parameter. It gives an interval estimate, and it is not focused on the concept of hypothesis rejection.
Information provided	Provides information about the statistical significance of an effect or relationship. It helps determine whether an observed effect is likely due to a real phenomenon or could have occurred by chance.	Provides a range of values within which the true population parameter is likely to fall. It gives a sense of the precision of the estimate.
Statement of findings	Conclusions from a hypothesis test typically involve rejecting or not rejecting the null hypothesis based on the	Conclusions from a confidence interval involve stating the range of values within which the population parameter is likely to lie

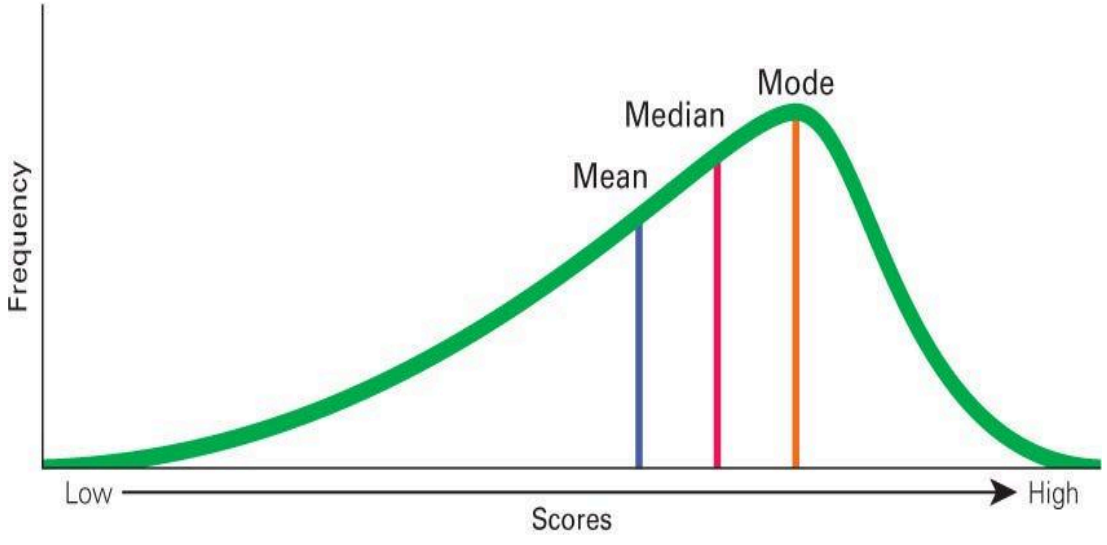
	evidence observed in the sample.	with a certain level of confidence.
Use of null hypothesis	In hypothesis testing, the null hypothesis is explicitly stated and is tested against an alternative hypothesis.	There is no explicit null hypothesis in the construction of a confidence interval.

Q36. What is the left-skewed distribution and the right-skewed distribution?

Skewness is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images.
A distribution can have right (or positive), left (or negative), or zero skewness. A right-skewed distribution is longer on the right side of its peak, and a left-skewed distribution is longer on the left side of its peak.

Left-Skewed distribution (negative skew):

- A left-skewed distribution is longer on the left side of its peak than on its right. In other words, a left-skewed distribution has a long tail on its left side. Left skew is also referred to as negative skew.
- The distribution is left-skewed because it's longer on the left side of its peak.



(b) Left-skewed distribution

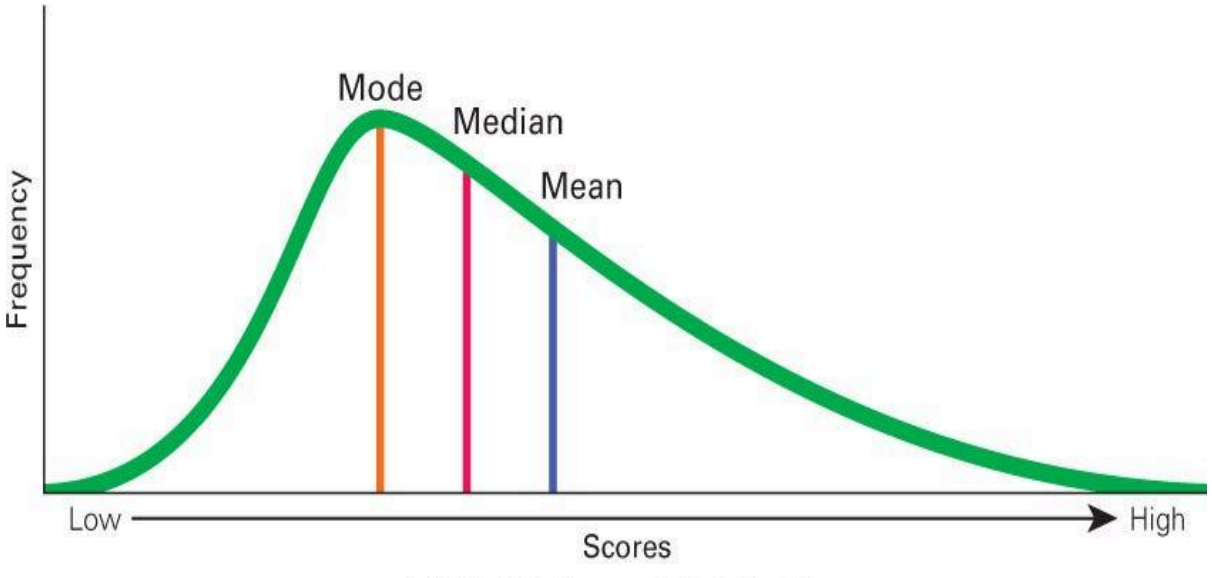
- The majority of the data points are concentrated on the right side, and the distribution stretches out to the left.

- The mean is typically less than the median in a left-skewed distribution because the presence of the longer left tail pulls the mean in that direction.

Left skew: mean < median

Right-Skewed distribution (Positive skew):

- A right-skewed distribution is longer on the right side of its peak than on its left. Right skew is also referred to as positive skew.
- You can think of skewness in terms of tails. A tail is a long, tapering end of a distribution. It indicates that there are observations at one of the extreme ends of the distribution, but that they're relatively infrequent. A right-skewed distribution has a long tail on its right side.



- The majority of the data points are concentrated on the left side, and the distribution stretches out to the right.
- The mean is typically greater than the median in a right-skewed distribution because the presence of the longer right tail pulls the mean in that direction.
- The mean of a right-skewed distribution is almost always greater than its median. That's because extreme values (the values in the tail) affect the mean more than the median.

Right skew: mean > median

These skewness characteristics provide information about the direction and degree of asymmetry in a dataset. It's important to note that skewness is just one aspect of a distribution's shape, and it should be interpreted in conjunction with other statistical measures and domain knowledge.

Q37. What is Bessel's correction?

Bessel's correction is a statistical adjustment made to correct bias in the estimation of the population variance and covariance based on a sample of observations. The correction is named after Friedrich Bessel, a German mathematician and astronomer who introduced it in the early 19th century.

The sample variance (s^2) and sample covariance (sxy) formulas without Bessel's correction are as follows:

1. **Sample Variance (without Bessel's correction):**
   $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$
2. **Sample Covariance (without Bessel's correction):**
   $$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Here, n is the number of observations, xi and yi are individual data points, $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively.

Bessel's correction involves dividing the sum of squared deviations by $(n-1)$ instead of n in the formulas above. The corrected sample variance (s^2 with Bessel's correction) and corrected sample covariance (sxy with Bessel's correction) are given by:

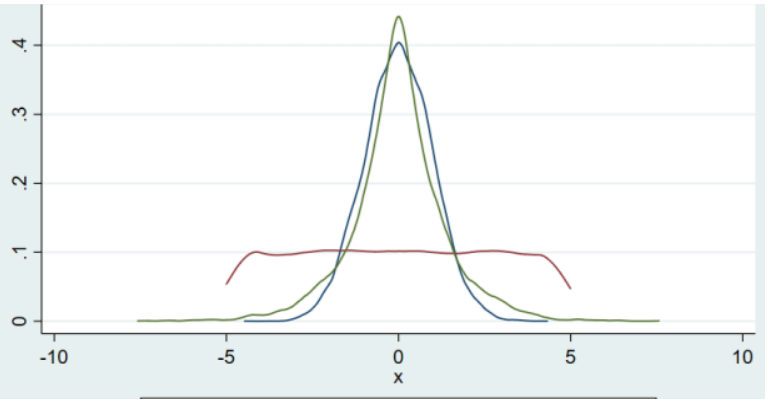
1. **Sample Variance (with Bessel's correction):**
   $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$
2. **Sample Covariance (with Bessel's correction):**
   $$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The reason for using *n*−1 instead of *n* is to correct the bias in the estimation of the population variance and covariance when working with a sample rather than the entire population. This correction provides a more accurate and unbiased estimate of the population parameters based on limited sample data.

Q38.What is Kurtosis?

Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution. Positive kurtosis indicates heavier tails and a more peaked distribution, while negative kurtosis suggests lighter tails and a flatter distribution. Kurtosis helps in analyzing the characteristics and outliers of a dataset.

The measure of Kurtosis refers to the tailedness of a distribution. Tailedness refers to how often the outliers occur. Peakedness in a data distribution is **the degree to which data values are concentrated around the mean**. Datasets with high kurtosis tend to have a distinct peak near the mean, decline rapidly, and have heavy tails. Datasets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.



The formula for sample kurtosis is often expressed as:

$$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)^2} - 3$$

Here:

- n is the number of observations in the sample,
- X_i is each individual observation,
- $\bar{X}$ is the sample mean.

The term $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4$ measures the fourth moment, and

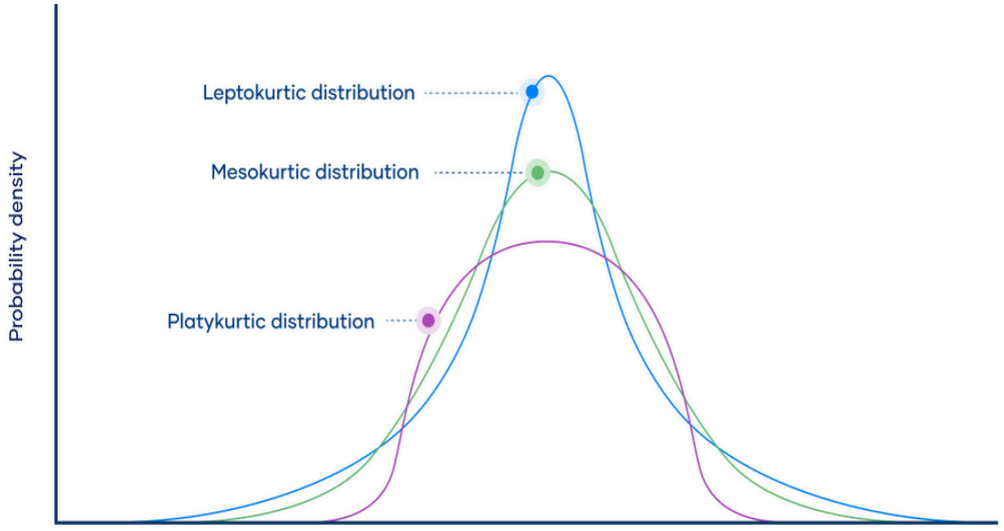
Excess Kurtosis

In statistics and probability theory, researchers use excess kurtosis to compare the kurtosis coefficient with that of a normal distribution. Since normal distributions have a kurtosis of 3, excess kurtosis is calculated by subtracting kurtosis by 3.

Excess kurtosis = Kurt – 3

Kurtosis can be classified into three main types:

- ***Leptokurtic (Kurtosis > 3)***
  Leptokurtic has very long and thick tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. Extremely positive kurtosis indicates a distribution where more numbers are located in the tails of the distribution instead of around the mean.



- ***Platykurtic (Kurtosis < 3)***
  Platykurtic having a thin tail and stretched around the center means most data points are present in high proximity to the mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

- ***Mesokurtic (Kurtosis = 3)***
  Mesokurtic is the same as the normal distribution, which means kurtosis is near 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

Q39. What is the probability of throwing two fair dice when the sum is 5 and 8?

When rolling two fair six-sided dice, each die has numbers from 1 to 6. The possible outcomes for the sum of the two dice range from 2 to 12.

Let's consider the cases where the sum is 5 and 8:

Sum of 5
Possible combinations: ***(1, 4), (2, 3), (3, 2), (4, 1)***
There are 4 favourable outcomes.

Sum of 8
Possible combinations: ***(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)***
There are 5 favourable outcomes.

The total number of possible outcomes when rolling two dice is 6×6 = 36(since each die has 6 faces).

Now, we can calculate the probability for each case:
Probability of getting a sum of 5:

$$\frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}} = \frac{4}{36}$$

Probability of getting a sum of 8:

$$\frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}} = \frac{5}{36}$$

These probabilities are based on the assumption that each outcome is equally likely when rolling fair dice. Therefore, the probability of getting a sum of 5 is 4/36 and the probability of getting a sum of 8 is 5/36.

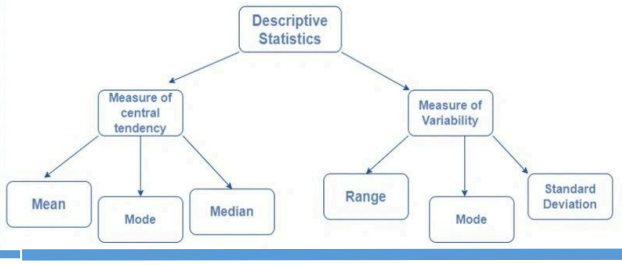
Q40. What is the difference between Descriptive and Inferential Statistics?

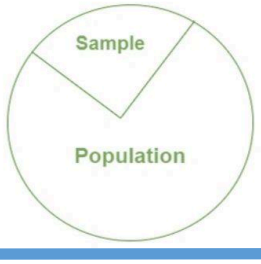
Descriptive statistics and inferential statistics are two branches of statistics that serve different purposes in analyzing and interpreting data.

Descriptive statistics: Descriptive statistics involves the collection, organization, summarization, and presentation of data to provide a clear and meaningful description of its main features.

Inferential statistics: Inferential statistics is the branch of statistics that deals with making inferences, predictions, and generalizations about a population based on a sample of data drawn from that population.

DESCRIPTIVE	INFERENTIAL
It is the analysis of data that helps to describe, show and summarize data under study	It is the analysis of random sample of data taken from a population to describe and make inference about the population
Organize, analyze and present data in a meaningful way	Compares, test and predicts data
It is used to describe a situation	It is used to explain the chance of occurrence of an event
It explain already known data and limited to a sample or population having small size	It attempts to reach the conclusion about the population
Types: Measure of central tendency & Measure of variability	Types: Estimation of parameters & Testing of hypothesis
Results are shown with help of charts, graphs, tables etc.	Results are shown with help of probability scores





Both types of statistics are essential in the field of data analysis and research. Descriptive statistics provide a foundation for understanding and interpreting data, while inferential statistics enable researchers to draw conclusions and make predictions about populations based on limited sample data.

Q41. 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 formula for calculating the raw score (X) from a z-score (Z) score is given by:

X=Mean+(Z×Standard Deviation)

In this case,

The mean (μ) is 160,

The standard deviation (σ) is 15,

And the z-score (Z) is 1.20. Substituting these values into the formula:

X=160+(1.20×15)

X=160+18

X=178

Therefore, Jeremy's score on the test would be 178.

Q42.What is the meaning of degrees of freedom (DF) in statistics?

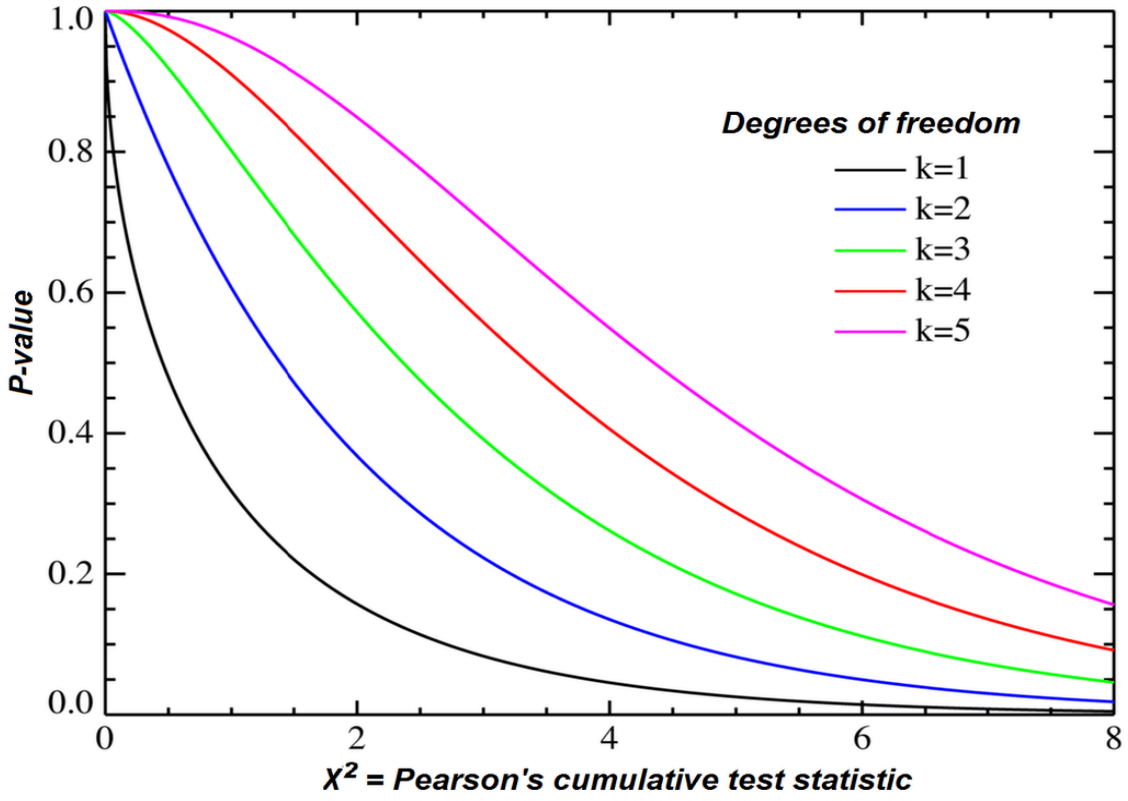
Degrees of freedom (DF) is a concept used in statistics to describe the amount of freedom or variability available in the estimation of a parameter. The specific meaning of degrees of freedom can vary depending on the context in which it is used.

Here are a few common scenarios where degrees of freedom are relevant:

- ***T-test and confidence intervals***: In the context of t-tests or confidence intervals for the mean, degrees of freedom represent the number of values in the final calculation of a statistic that are free to vary. For example, in a t-test comparing the means of two groups, the degrees of freedom would be related to the sample sizes of the two groups.

- ***Chi square test:*** In chi-square tests of independence or goodness of fit, degrees of freedom represent the number of categories in the data that are free to vary. It depends on the number of categories in the variable minus 1.
- ***ANOVA (Analysis of variance):*** In the context of analysis of variance, degrees of freedom are used to describe the variability between groups and within groups. There are degrees of freedom associated with the between-group variance and within-group variance.

- ***Regression analysis:*** In multiple regression analysis, degrees of freedom can be associated with the number of predictors in the model and the sample size. Degrees of freedom are used in hypothesis testing for the overall significance of the regression model and the significance of individual regression coefficients.

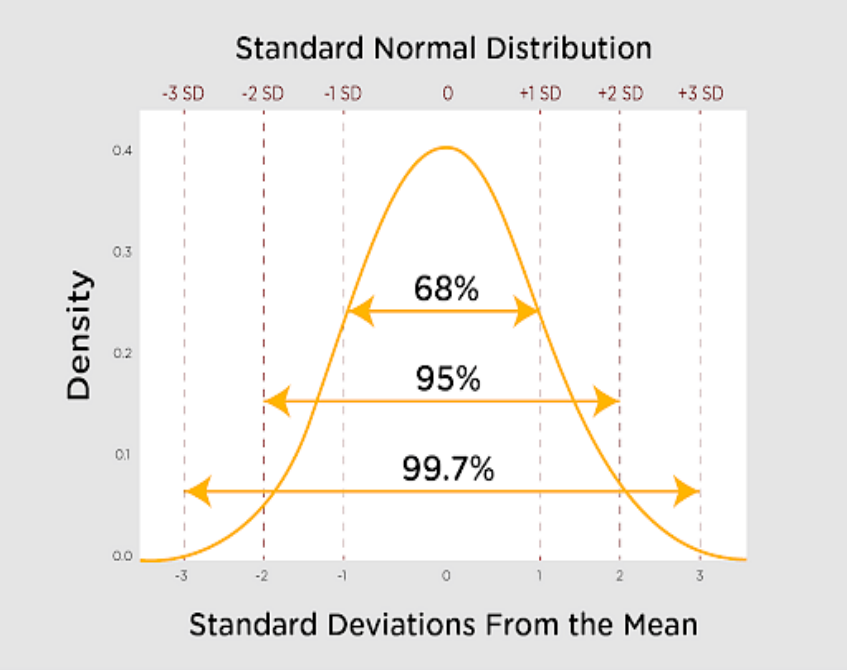


In general, degrees of freedom are a measure of the amount of data that are available to estimate parameters in a statistical model. The concept is crucial in statistical hypothesis testing, where it is used to determine critical values for test statistics and to assess the precision of parameter estimates.

Understanding degrees of freedom is fundamental for interpreting and conducting various statistical tests.

Q43. What is the empirical rule in Statistics?

The empirical rule, also known as the three-sigma rule or the 68-95-99.7 rule, is a statistical rule that states that almost all observed data for a normal distribution will fall within three standard deviations (denoted by σ) of the mean or average (denoted by μ). This rule is based on the properties of the normal distribution, also known as the bell curve or Gaussian distribution.



The normal distribution is associated with the 68-95-99.7 rule

68% of the data is within 1 standard deviation (σ) of the mean (μ).

95% of the data is within 2 standard deviations (σ) of the mean (μ).

99.7% of the data is within 3 standard deviations (σ) of the mean (μ).

The formula for Empirical Rule is:

$$\mu \pm m\sigma$$

μ = Mean, σ = Standard deviation. m = Multiplier

Conclusion

Empirical Rule is a statistical concept that aids in showing the probability of observations and is particularly useful when approximating a large population. It's important to remember that these are only estimates. There is always the possibility of outliers who do not fit into the distribution. As a result, the

findings are inaccurate, and you should exercise caution when acting on the forecast.

Q44. What is the relationship between sample size and power in hypothesis testing?

The relationship between sample size and power in hypothesis testing is a crucial consideration in experimental design and statistical analysis.

A few key points regarding the relationship between sample size and power are:

- ***Increasing sample size increases power***: Generally, as the sample size increases, the statistical power of a test tends to increase. Larger sample sizes provide more information and reduce the variability in the data, making it easier to detect a true effect if it exists. This relationship is particularly important in detecting small or subtle effects.
- ***Power analysis:*** Before conducting a study, researchers often perform a power analysis to determine the sample size required to achieve a desired level of statistical power. Power analysis takes into account factors such as the effect size, significance level, and variability in the data.
- ***Trade- off between and practical constraints:*** While increasing the sample size can enhance power, there is often a practical limit imposed by factors such as time, cost, and feasibility.
- ***Effect size matters:*** The relationship between sample size and power is influenced by the size of the effect being investigated. Smaller effects may require larger sample sizes to achieve the same level of power. Researchers should consider the meaningfulness of the effect size in the context of the study.
- ***Statistical significance vs practical significance:*** A study with a large sample size may achieve statistical significance for even small effects, but it's essential to assess whether the observed effect is practically significant and meaningful in the real-world context.

Conclusion

The relationship between sample size and power is positive; increasing the sample size generally increases the power of a statistical test. However, researchers need to carefully consider factors such as effect size, practical constraints, and the balance between statistical and practical significance when determining the appropriate sample size for a study.

Q45. How does increasing the confidence level affect the width of a confidence interval?

The confidence level and the width of a confidence interval are inversely related. As you increase the confidence level, the width of the confidence interval also increases, and vice versa. The confidence level represents the probability that the true parameter (e.g., a population mean or proportion) lies within the calculated confidence interval. Commonly used confidence levels include 90%, 95%, and 99%, among others.

The width of a confidence interval is influenced by two main factors: the standard deviation of the data and the sample size. The formula for a confidence interval is:

Confidence Interval=Point Estimate±(Critical Value×Standard Error)

The critical value is determined by the desired confidence level, and the standard error depends on the standard deviation and the sample size. As you increase the confidence level, the critical value becomes larger, leading to a wider interval. The desired level of confidence will affect the width of the confidence interval. Given the same data, if we want to have more confidence that the DGP falls within a specified range, we will have to make our confidence interval wider.

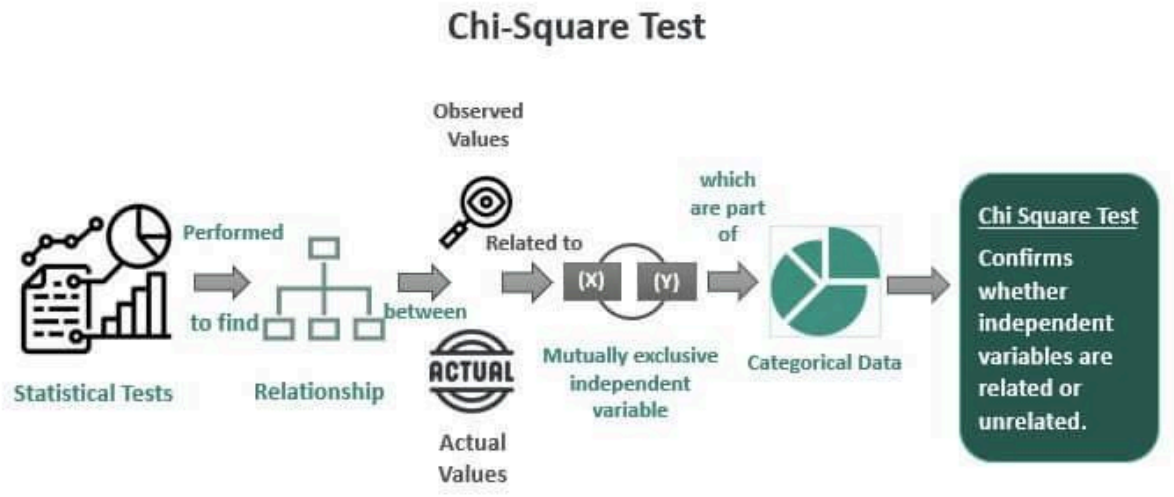
Consider this extreme example: if we want to be 100% confident that the true value of β1 falls within the confidence interval, we would need the interval to go from negative infinity to positive infinity – as wide as a confidence interval possibly could be! That's the only way we could have 100% confidence. If we want just 95% confidence, we could make the interval narrower (whew!). And if we want even less confidence (e.g., 90% or 80%), the interval could get even narrower.

In summary, higher confidence levels require larger critical values, resulting in wider confidence intervals. Lower confidence levels involve smaller critical

values, leading to narrower intervals. Keep in mind that a wider interval implies greater uncertainty about the true parameter value but provides a higher level of confidence in capturing that value.

Q46. What is a Chi-Square test?

A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables. For example, a meal delivery firm in India wants to investigate the link between gender, geography, and people's food preferences.



It is used to calculate the difference between two categorical variables, which are:

- ***As a result of chance or***
- ***Because of the relationship***

Formula For Chi-Square Test

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

Where

c = Degrees of freedom

O = Observed Value

E = Expected Value

There are different types of Chi-Square tests, but the two main ones are the Chi-Square test for independence and the Chi-Square goodness-of-fit test:

Chi-Square test for independence:

This test is used when you want to determine if there is a significant association between two categorical variables.
It involves setting up a contingency table to compare the observed frequencies of the joint distribution of the variables with the frequencies that would be expected if the variables were independent.
The test statistic follows a Chi-Square distribution, and the calculation involves comparing observed and expected frequencies.

Chi-Square goodness of fit test:

This test is used when you want to assess whether an observed frequency distribution differs from a theoretical (expected) distribution.
It involves comparing the observed frequencies in different categories with the frequencies that would be expected if the data followed a particular distribution.
The test statistic is calculated based on the differences between observed and expected frequencies.

Conclusion

A chi-square distribution is followed by very few real-world observations. The objective of chi-square distributions is to test hypotheses, not to describe real-world distributions. In contrast, most other commonly used distributions, such as normal and Poisson distributions, may explain important things like baby birth weights or illness cases per year. Because of its close resemblance to the conventional normal distribution, chi-square distributions are excellent for hypothesis testing. Many essential statistical tests rely on the conventional normal distribution.

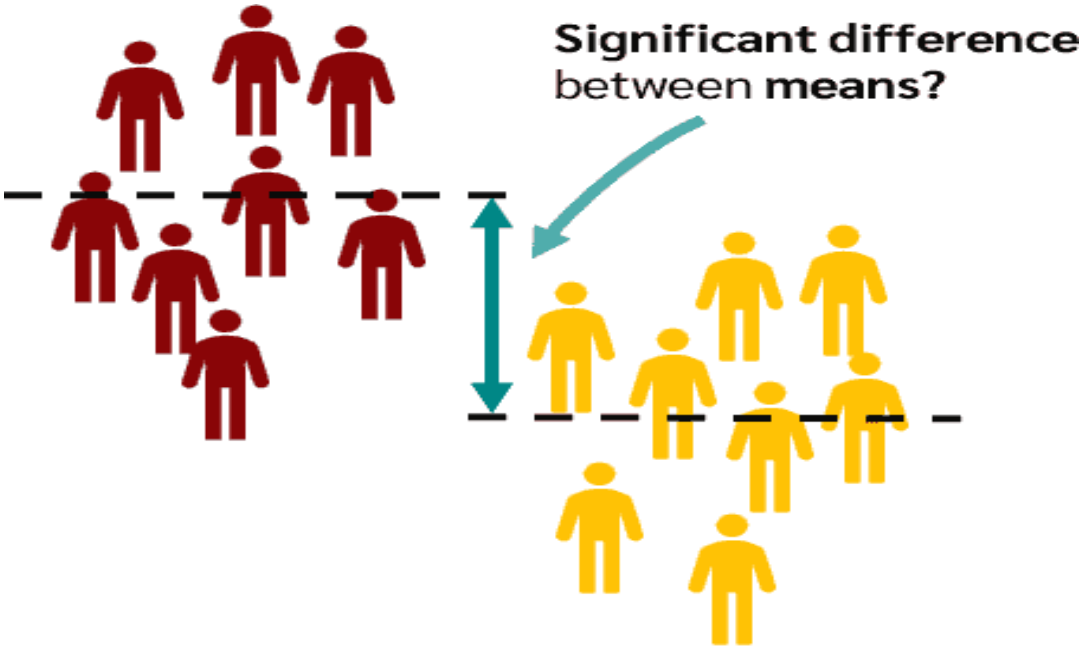
The chi-square test, for starters, is extremely sensitive to sample size. Even insignificant relationships can appear statistically significant when a large enough sample is used. Keep in mind that "statistically significant" does not always imply "meaningful" when using the chi-square test.

Be mindful that the chi-square can only determine whether two variables are related. It does not necessarily follow that one variable has a causal relationship with the other. It would require a more detailed analysis to establish causality.

Q47. What is a t-test?

A t-test is a statistical method used to determine if there is a significant difference between the means of two groups. It is particularly useful when dealing with small sample sizes and assumes that the data is approximately normally distributed. The t-test is widely used in various fields, including biology, psychology, business, and other sciences.

The two groups could be, for example, patients who received *drug A* once and *drug B* once, and you want to know if there is a difference in blood pressure between these two groups.



The t-test is based on the t-distribution, which is a mathematical distribution similar to the normal distribution but with heavier tails. The test calculates a t-statistic, which measures the difference between the means of the two groups in terms of the standard error of the difference. The larger the t-statistic, the more likely it is that the difference between the groups' means is not due to random chance.

There are several types of t-tests, but the two most common ones are:

The Independent Samples t-test and

The Paired Samples t-test.

The Independent Samples t-test:

- This test is used when comparing the means of two independent groups to determine if there is a significant difference between them.
- The independent groups could be, for example, two different treatment groups, a treatment group and a control group, or any other two unrelated groups.
- The test compares the means of the two groups while considering the variability within each group.
- The formula for the t-test statistic in the independent samples t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where,

- $\bar{X}1$ and $\bar{X}2$ are the sample means of the two groups,
- $s1$ and $2s2$ are the sample standard deviations of the two groups,
- $n1$ and $n2$ are the sample sizes of the two groups.

Paired samples t-test:

- This test is used when comparing the means of two related groups, such as repeated measurements on the same subjects.
- It assesses whether the mean difference between paired observations is significantly different from zero.
- The test is often applied in situations where the same subjects are measured before and after a treatment.
- The formula for the t-test statistic in the paired samples t-test is:

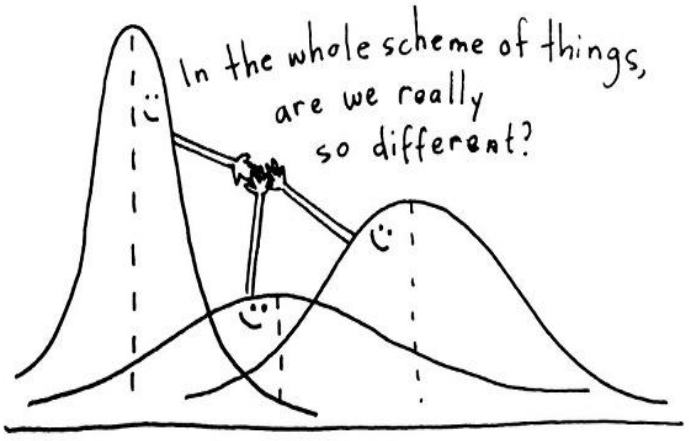
$$t = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}}$$

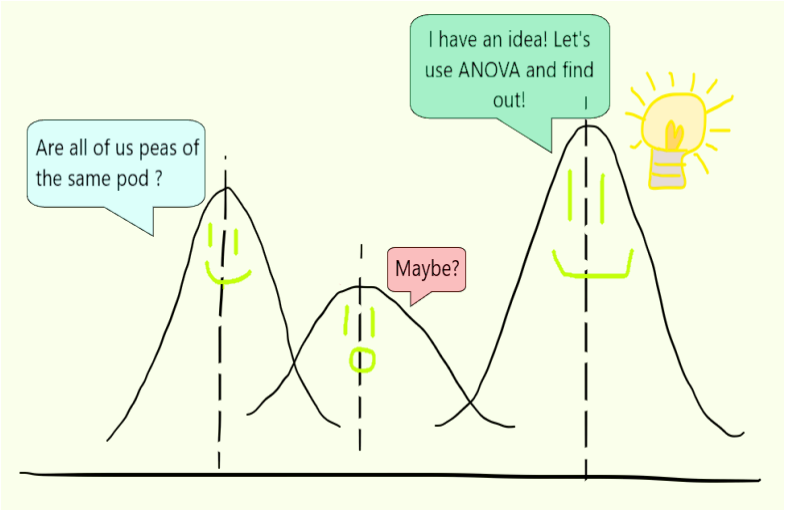
Where,

- ¯D is the mean of the paired differences,
- *sD* is the standard deviation of the paired differences,
- *n* is the number of pairs.

Q48. What is the ANOVA test?

ANOVA stands for Analysis of Variance. It is a statistical method used to analyze the differences between the means of two or more groups or treatments. It is often used to determine whether there are any statistically significant differences between the means of different groups. ANOVA compares the variation between group means to the variation within the groups. If the variation between group means is significantly larger than the variation within groups, it suggests a significant difference between the means of the groups.





Example

Consider a scenario where we have three medical treatments for patients with similar diseases. Once we have the test results, one approach is to assume that the treatment which took the least time to cure the patients is the best among them. What if some of these patients had already been partially cured, or if any other medication was already working on them?

In order to make a confident and reliable decision, we will need evidence to support our approach. This is where the concept of ANOVA comes into play. A common approach to figuring out a reliable treatment method would be to analyze the days the patients took to be cured. We can use a statistical technique to compare these three treatment samples and depict how different these samples are from one another. Such a technique, which compares the samples based on their means, is called ANOVA. ANOVA checks the impact of

one or more factors by comparing the means of different samples. We can use ANOVA to prove/disprove whether all the medication treatments were equally effective.

There are different types of ANOVA, but the most common ones are:

- One way ANOVA: This is used when there is one independent variable with more than two levels or groups. It tests whether there are any significant differences between the means of the groups.
- Two way ANOVA: This involves two independent variables. It is used to explore the interaction effect between these variables on the dependent variable.
- Repeated measures ANOVA: This is an extension of one-way or two-way ANOVA used when the same subjects are used for each treatment (e.g., repeated measurements taken on the same individuals).

BASIS FOR COMPARISON	ONE WAY ANOVA	TWO WAY ANOVA
Meaning	One way ANOVA is a hypothesis test, used to test the equality of three of more population means simultaneously using variance.	Two way ANOVA is a statistical technique wherein, the interaction between factors, influencing variable can be studied.
Independent Variable	One	Two
Compares	Three or more levels of one factor.	Effect of multiple level of two factors.
Number of Observation	Need not to be same in each group.	Need to be equal in each group.
Design of experiments	Need to satisfy only two principles.	All three principles needs to be satisfied.

Conclusion

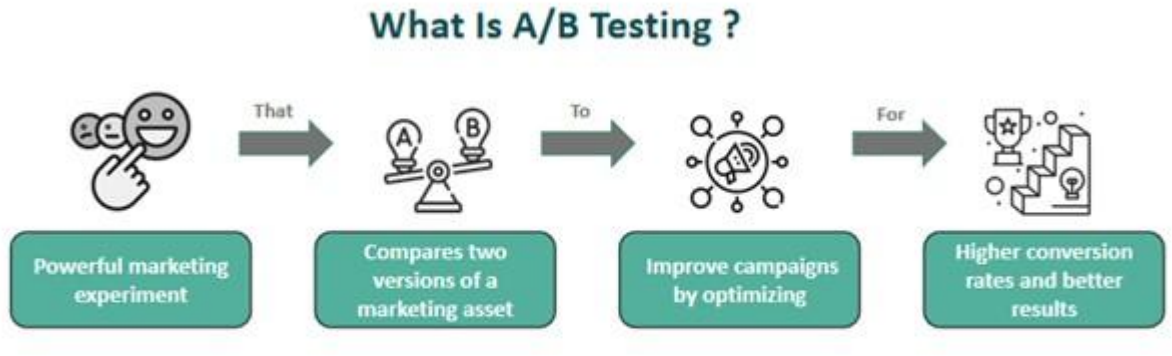
The hypothesis in ANOVA is that the means of the groups are equal. If the observed differences between the group means are larger than what would be expected by random chance, the ANOVA test will indicate that there is a significant difference.

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.

Q49. How is hypothesis testing utilised in A/B testing for marketing campaigns?

Hypothesis testing plays a crucial role in A/B testing for marketing campaigns. A/B testing, also known as split testing, is a method used to compare two versions (A and B) of a marketing campaign or webpage to determine which one performs better in terms of a predefined metric. This could be click-through rates, conversion rates, revenue generated, or any other key performance indicator (KPI) relevant to the marketing goal.



A/B testing is defined as a controlled experiment involving two versions of a marketing asset getting tested simultaneously to determine better business driving metrics, impacting visitors the most. Then, these two versions are exposed to different portions of the audience to identify and measure the effective version that can achieve the objective.

It is done by deploying a tracking mechanism for metrics measurement and performance determination up to a specific period. Finally, the results are fed into statistical tools and techniques for drawing conclusions and selecting the winning variation. It is the most accessible and versatile marketing tool that any business size can apply. It helps firms achieve a higher **conversion rate** of visitors into customers and enhances their marketing efforts. It has wide use in direct mailers, advertisements, landing pages, websites, and email campaigns. Marketers use it to test various aspects of their **marketing campaign**, like promotional offers, design, messaging, and **pricing strategies**.

Example

Spotify, the music streaming platform, has introduced Confidence, a tool for software development teams. This tool empowers teams to efficiently set up, run, coordinate, and analyze user tests, including A/B testing and more complex scenarios. Currently, Confidence is in a private beta phase, offering access to a select group of users.
In terms of A/B testing, Spotify's blog post underscores the wealth of experience of its scientists and engineers in perfecting product testing methods over the years. It includes conducting simultaneous A/B tests and deploying AI recommendation systems across various platforms.

Confidence offers three real-world access points: a managed service, a backstage plugin, and API integration. Users eager to embrace Confidence can join a waitlist, though a specific product release date remains undisclosed, mirroring the anticipation that often surrounds such A/B testing tools.

Here's how hypothesis testing is typically utilized in the A/B testing process:

1. Formulating hypothesis:

- ***Null hypothesis (H_0):*** There is no significant difference between the two versions (A and B).
- ***Alternate hypothesis (H_a):*** There is a significant difference between the two versions, and one performs better than the other.

For example:

- H_0: The click-through rates of version A and version B are equal.
- H_a: The click-through rate of version A is different from the click-through rate of version B.

2. Selecting a significance level: The significance level (often denoted as alpha, α) is the threshold at which you decide whether to reject the null hypothesis. Commonly used values are 0.05 or 0.01.

3. Collecting data: Implement versions A and B of the marketing campaign and collect relevant data on the chosen metric(s).

4. Performing test: Use statistical methods, often t-tests or z-tests, to analyze the collected data. This helps determine if the observed differences between version A and version B are statistically significant.

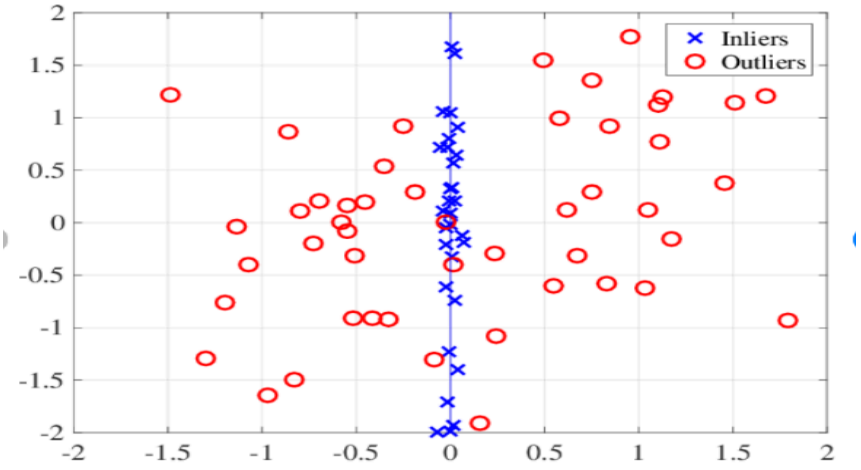
It's important to note that A/B testing should be conducted with proper randomization, control groups, and a sufficiently large sample size to ensure the reliability of the results. Additionally, multiple testing corrections may be

necessary if you are running multiple comparisons simultaneously to avoid inflated Type I error rates.

Q50. What is an inlier?

An inlier is a data point that lies within a specific pattern or cluster in a dataset. Inliers are typically close to the central tendency of the data and are consistent with the general trend or behaviour of the majority of the data points. Inlier detection is often used in the context of outlier detection. Because inliers are difficult to distinguish from good data values they are sometimes difficult to find and correct. A simple example of an inlier might be a value in a record reported in the wrong units, say degrees Fahrenheit instead of degrees Celsius.

By this definition the inlier has two aspects: (1) it is in the interior of the relevant distribution of values; and (2) it is an erroneous value.



Dealing with inliers (which really generally involves *not* dealing with them):

Unless you have a source of external information indicating measurement error, it is essentially impossible to identify "inliers". By definition, these are data points that are in the "interior" of the distribution, where most of the other data occurs. Hence, it is not detected by tests that look for data that is an "aberration" from the other data points. (In some cases you can detect "inliers" that seem to be in the interior of a distribution, but are actually "outliers" when taken with respect to a more complex representation of the distribution. In this case the point is actually an outlier, but it only looks like it is in the interior of the distribution when you are using a crude distributional approximation.)

In some rare cases you might have an external source of information that identifies a subset of your data as being subject to measurement error (e.g., if you are conducting a large survey and you find out that one of your surveyors was just making up their data). In this case, any data points in that subset that are in the interior of the distribution are "inliers" and are known via external information to be subject to measurement error. In this case you would generally remove all the data known to be erroneous, even if some of it is "inliers" that are in the interior of the distribution where you would expect it to be. The point here is that a data point can be erroneous even if it is not in the tails of the distribution.

Conclusion

Inlier and outlier analysis is crucial in various fields, including statistics, machine learning, and data analysis. Detecting outliers can help identify errors, anomalies, or interesting patterns in the data. Conversely, recognizing inliers is important for understanding the typical behaviour or characteristics of a dataset. Different techniques, such as statistical methods or machine learning algorithms, can be employed to identify and handle inliers and outliers in a dataset.