1. What is the Central Limit Theorem and why is it important?

“Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can’t obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample. The Central Limit Theorem addresses this question exactly.”

2. What is sampling? How many sampling methods do you know?

“Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined.”

3. What is the difference between type I vs type II error?

“A type I error occurs when the null hypothesis is true, but is rejected. A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected.”

4. What is linear regression? What do the terms p-value, coefficient, and r-squared value mean? What is the significance of each of these components?

A linear regression is a good tool for quick predictive analysis: for example, the price of a house depends on a myriad of factors, such as its size or its location. In order to see the relationship between these variables, we need to build a linear regression, which predicts the line of best fit between them and can help conclude whether or not these two factors have a positive or negative relationship.

5. What are the assumptions required for linear regression?

There are four major assumptions: 1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data, 2. The errors or residuals of the data are normally distributed and independent from each other, 3. There is minimal multicollinearity between explanatory variables, and 4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

6. What is a statistical interaction?

” Basically, an interaction is when the effect of one factor (input variable) on the dependent variable (output variable) differs among levels of another factor.”

7. What is selection bias?

“Selection (or ‘sampling’) bias occurs in an ‘active,’ sense when the sample data that is gathered and prepared for modeling has characteristics that are not representative of the true, future population of cases the model will see. That is, active selection bias occurs when a subset of the data are systematically (i.e., non-randomly) excluded from analysis.”

8. What is an example of a data set with a non-Gaussian distribution?

“The Gaussian distribution is part of the Exponential family of distributions, but there are a lot more of them, with the same sort of ease of use, in many cases, and if the person doing the machine learning has a solid grounding in statistics, they can be utilized where appropriate.”

9. What is the Binomial Probability Formula?

“The binomial distribution consists of the probabilities of each of the possible numbers of successes on N trials for independent events that each have a probability of π (the Greek letter pi) of occurring.”

10. What is the difference between “long” and “wide” format data?

In the **wide-format**, a subject’s repeated responses will be in a single row, and each response is in a separate column. In the **long-format**, each row is a one-time point per subject. You can recognize data in wide format by the fact that columns generally represent groups.

Name	Height	Weight
John	160	67
Christopher	182	78

Figure: *Wide Format*

Name	Attribute	Value
John	Height	160
John	Weight	67
Christopher	Height	182
Christopher	Weight	78

Figure: *Long Format*

11. What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up.

However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

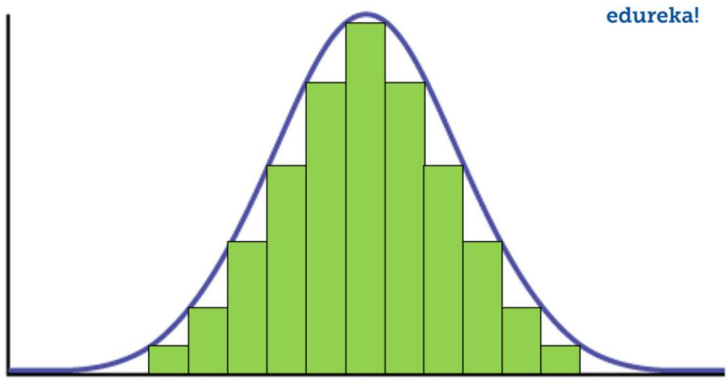


Figure: *Normal distribution in a bell curve*

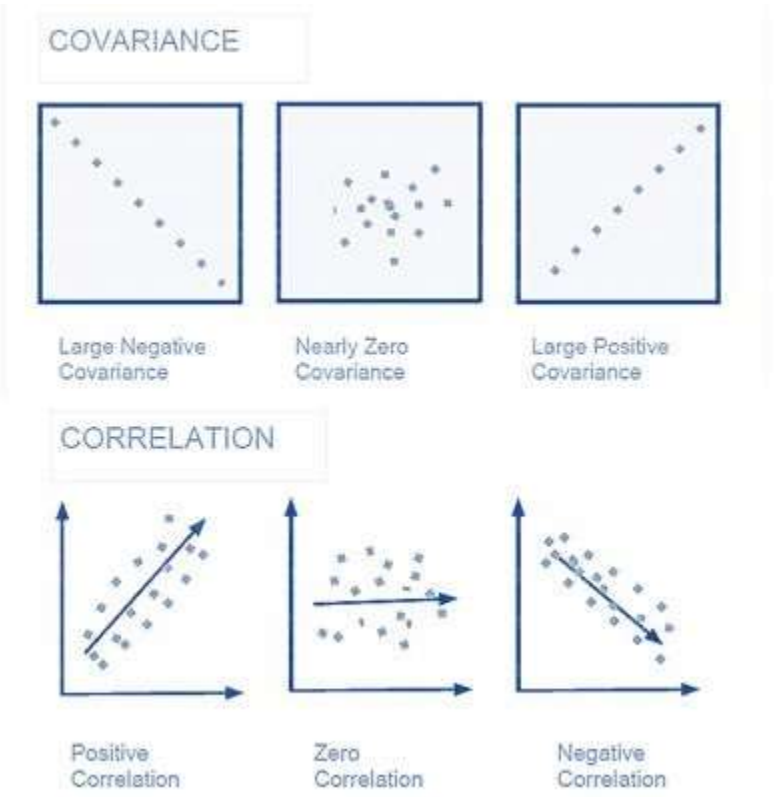
The random variables are distributed in the form of a symmetrical, bell-shaped curve.

Properties of Normal Distribution are as follows;

1. Unimodal -one mode
2. Symmetrical -left and right halves are mirror images
3. Bell-shaped -maximum height (mode) at the mean
4. Mean, Mode, and Median are all located in the center
5. Asymptotic

12. What is correlation and covariance in statistics?

Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both Correlation and Covariance establish the relationship and also measure the dependency between two random variables. Though the work is similar between these two in mathematical terms, they are different from each other.



Correlation: Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.

Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

13. What is the difference between Point Estimates and Confidence Interval?

Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.

A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by 1 — alpha, where alpha is the level of significance.

14. What is the goal of A/B Testing?

It is a hypothesis testing for a randomized experiment with two variables A and B.

The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business. It can be used to test everything from website copy to sales emails to search ad. An example of this could be identifying the click-through rate for a banner ad.

15. What is p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called the Null Hypothesis.

Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,

High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

16. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?

Probability of not seeing any shooting star in 15 minutes is

= 1–P(Seeing one shooting star)
= 1 – 0.2 = 0.8

Probability of not seeing any shooting star in the period of one hour

= (0.8) ^ 4 = 0.4096

Probability of seeing at least one shooting star in the one hour

= 1–P(Not seeingany star)
= 1 – 0.4096 = 0.5904
Probability of not seeing any shooting star in the period of one hour

= (0.8) ^ 4 = 0.4096

Probability of seeing at least one shooting star in the one hour

= 1 - P (Not seeing any star)
= 1-0.4096
=0.5904

17. How can you generate a random number between 1 – 7 with only a die?

- Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes.
- To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one.
- A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice.
- All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

18. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

In the case of two children, there are 4 equally likely possibilities

BB, BG, GB and GG;

where **B** = Boy and **G** = Girl and the first letter denotes the first child.

From the question, we can exclude the first case of BB. Thus from the remaining 3 possibilities of **BG**, **GB** & **BB**, we have to find the probability of the case with two girls.

Thus, P(Having two girls given one girl) = **1 / 3**

19. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

There are two ways of choosing the coin. One is to pick a fair coin and the other is to pick the one with two heads.

Probability of selecting fair coin = 999/1000 = 0.999
Probability of selecting unfair coin = 1/1000 = 0.001

Selecting 10 heads in a row = Selecting fair coin * Getting 10 heads + Selecting an unfair coin

P(A) = 0.999 *(1/2)^5 = 0.999 * (1/1024) = 0.000976

P(B) = 0.001 *1 = 0.001

P(A / A + B) = 0.000976 / (0.000976 + 0.001) = 0.4939

P(B / A + B) = 0.001 / 0.001976 = 0.5061

Probability of selecting another head = P(A/A+B) * 0.5 + P(B/A+B) * 1 = 0.4939 * 0.5 + 0.5061 = 0.7531

Q 20. What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.). Sensitivity is nothing but “Predicted True events/ Total events”. True events here are the events which were true and model also predicted them as true.

Calculation of seasonality is pretty straightforward.
Seasonality = (True Positives) / (Positives in Actual Dependent Variable)

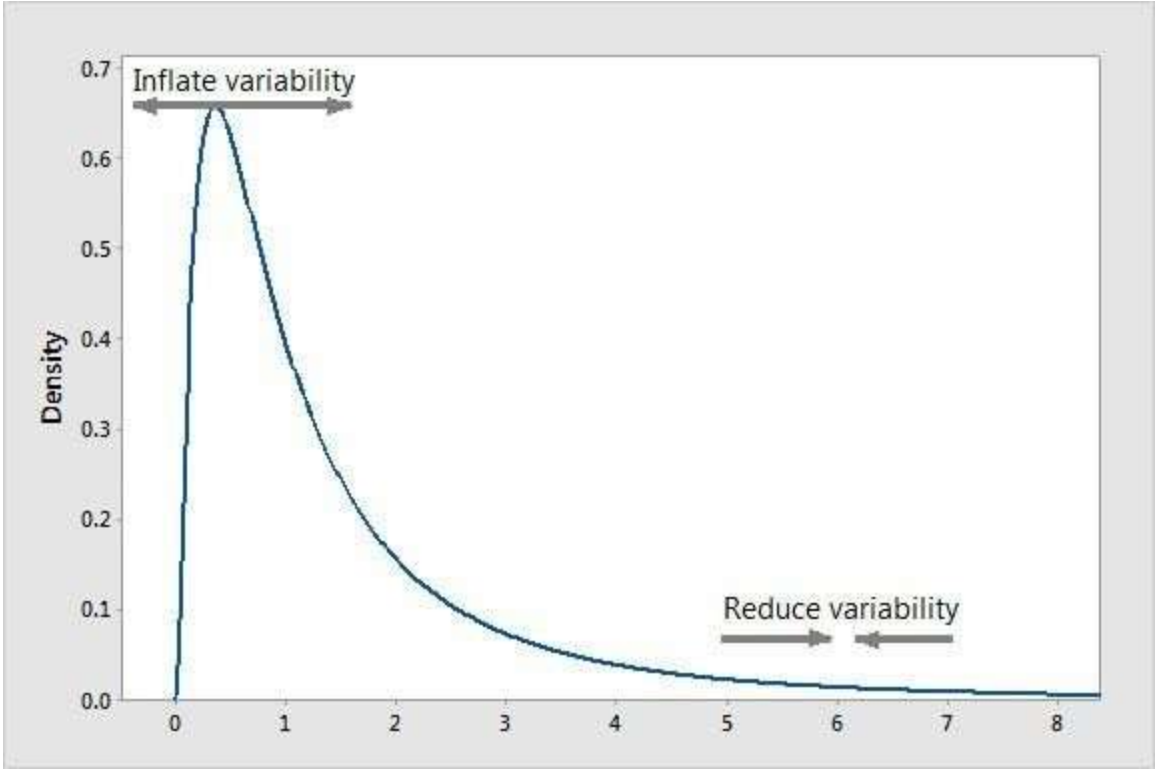
21. Why Is Re-sampling Done?

Resampling is done in any of these cases:

- Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points
- Substituting labels on data points when performing significance tests
- Validating models by using random subsets (bootstrapping, cross-validation)

22.What is a Box-Cox Transformation?

The dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow the skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.



A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box-Cox transformation is named after statisticians ***George Box*** and ***Sir David Roxbee Cox*** who collaborated on a 1964 paper and developed the technique.

23 You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?

This question has enough hints for you to start thinking! Since, the data is spread across median, let's assume it's a normal distribution. We know, in a normal distribution, ~68% of the data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

24. Is it possible capture the correlation between continuous and categorical variable? If yes, how?

Answer: Yes, we can use ANOVA (analysis of covariance) technique to capture association between continuous and categorical variables.

25. Y**ou are working on a classification problem. For validation purposes, you've randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. What went wrong?**

In case of classification problem, we should always use stratified sampling instead of random sampling. A random sampling doesn't takes into consideration the proportion of target classes. On the contrary, stratified sampling helps to maintain the distribution of target variable in the resultant distributed samples also.

26.What is Hypothesis Testing (Null and Alternate) ?

Hypothesis Testing - way for us to test the results of a survey or experiment, to see whether it have meaningful results.(basically **testing** whether your results are valid)

Null Hypothesis – when there is no statistical significance (importance) between the two variables in the hypothesis

Alternate Hypothesis – when there is statistical significance(importance) between the two variables in the hypothesis

27.What is one tail and two tail test ?

One tail - A statistical hypothesis test in which alternative hypothesis has only one end, is known as one tailed test.
Result ; - Greater or less than certain value.

two tail test - A statistical hypothesis test in which alternative hypothesis has only one end, is known as one tailed test.
Result ;- Greater or less than range of value

28. What is standard error ?
The standard error (SE) of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution[1] or an estimate of that standard deviation. If the parameter or the statistic is the mean, it is called the standard error of the mean.