Statistics Interview Questions

Question 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.”

Question 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.” There are two main types of Sampling techniques:Probability SamplingNon-Probability Sampling

Question 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.”

Question 4: What is linear regression?

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.

Question 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, and4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

Question 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.”.

Question 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.”

Question 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.”

Question 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.

The binomial distribution formula is:b(x; n, P) = nCx * Px * (1 – P)n – xWhere:

b = binomial probability

x = total number of “successes” (pass or fail, heads or tails, etc.)

P = probability of success on an individual trial

n = number of trials

Question 10: What is statistical power?

Wikipedia defines Statistical power or sensitivity of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true.

To put in another way, Statistical power is the likelihood that a study will detect an effect when the effect is present. The higher the statistical power, the less likely you are to make a Type II error (concluding there is no effect when, in fact, there is).

Question 11: Explain what resampling methods are and why they are useful. Also explain their limitations

.Classical statistical parametric tests compare observed statistics to theoretical sampling distributions. Resampling a data-driven, not theory-driven methodology which is based upon repeated sampling within the same sample.

Resampling refers to methods for doing one of theseEstimating the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping)Exchanging labels on data points when performing significance tests (permutation tests, also called exact tests, randomization tests, or re-randomization tests)Validating models by using random subsets (bootstrapping, cross validation)

Question 12: What is selection bias, why is it important and how can you avoid it?

Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample. For example, if a given sample of 100 test cases was made up of a 60/20/15/5 split of 4 classes which actually occurred in relatively equal numbers in the population, then a given model may make the false assumption that probability could be the determining predictive factor. Avoiding non-random samples is the best way to deal with bias; however, when this is impractical, techniques such as resampling, boosting, and weighting are strategies which can be introduced to help deal with the situation.

Question 13: 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.

Question 14: 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 curveThe random variables are distributed in the form of a symmetrical, bell-shaped curve.Properties of Normal Distribution are as follows;Unimodal -one modeSymmetrical -left and right halves are mirror imagesBell-shaped -maximum height (mode) at the meanMean, Mode, and Median are all located in the centerAsymptotic

Question 15: 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.

Question 16: 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.

Question 17: 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 adsAn example of this could be identifying the click-through rate for a banner ad.

Question 18: 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.

Question 19: 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 = 0.8Probability of not seeing any shooting star in the period of one hour= (0.8) ^ 4 = 0.4096Probability of seeing at least one shooting star in the one hour= 1 – P(Not seeing any star)

= 1 – 0.4096 = 0.5904

Question 20: 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.

Question 21: 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

Question 22: 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 = 0.001Selecting 10 heads in a row = Selecting fair coin * Getting 10 heads + Selecting an unfair coinP (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 = 0.5061Probability of selecting another head = P(A/A+B) * 0.5 + P(B/A+B) * 1 = 0.4939 * 0.5 + 0.5061 = 0.7531

Question 23: 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)

Question 24: 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 pointsSubstituting labels on data points when performing significance testsValidating models by using random subsets (bootstrapping, cross-validation)

Question 25: What are the differences between over-fitting and under-fitting?

In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data, so as to be able to make reliable predictions on general untrained data.

In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfitted, has poor predictive performance, as it overreacts to minor fluctuations in the training data.Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

Question 26: How to combat Overfitting and Underfitting?

To combat overfitting and underfitting, you can resample the data to estimate the model accuracy (k-fold cross-validation) and by having a validation dataset to evaluate the model.

Question 27: What is regularisation?

Why is it useful?

Regularisation is the process of adding tuning parameter to a model to induce smoothness in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1(Lasso) or L2(ridge). The model predictions should then minimize the loss function calculated on the regularized training set.

Q**uestion 28: What Is the Law of Large Numbers?**

It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of frequency-style thinking. It says that the sample means, the sample variance and the sample standard deviation converge to what they are trying to estimate.

Question 29: What Are Confounding Variables?

In statistics, a confounder is a variable that influences both the dependent variable and independent variable.For example, if you are researching whether a lack of exercise leads to weight gain,lack of exercise = independent variableweight gain = dependent variable.A

confounding variable here would be any other variable that affects both of these variables, such as the age of the subject.

Question 30: What Are the Types of Biases That Can Occur During Sampling?

Selection biasUnder coverage biasSurvivorship bias

Question 31: What is Survivorship Bias?

It is the logical error of focusing aspects that support surviving some process and casually overlooking those that did not work because of their lack of prominence. This can lead to wrong conclusions in numerous different means.

Q**uestion 32: What is selection Bias?**

Selection bias occurs when the sample obtained is not representative of the population intended to be analysed.

Question 33: Explain how a ROC curve works?

The ROC curve is a graphical representation of the contrast between true positive rates and false-positive rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity(true positive rate) and false-positive rate.

Question 34: What is TF/IDF vectorization?

TF–IDF is short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.The TF–IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Q**uestion 35: Why we generally use Softmax non-linearity function as last operation in-network?**

It is because it takes in a vector of real numbers and returns a probability distribution. Its definition is as follows. Let x be a vector of real numbers (positive, negative, whatever, there are no constraints).Then the i'th component of Softmax(x) is —

It should be clear that the output is a probability distribution: each element is non-negative and the sum over all components is 1.

Question 37: Python or R – Which one would you prefer for text analytics?

We will prefer Python because of the following reasons:Python would be the best option because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.R is more suitable for machine learning than just text analysis.Python performs faster for all types of text analytics.

Question 38: How does data cleaning plays a vital role in the analysis?

Data cleaning can help in the analysis because:Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.Data Cleaning helps to increase the accuracy of the model in machine learning.It is a cumbersome process

because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

Question 39: Differentiate between univariate, bivariate, and multivariate analysis.

Univariate analyses are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable, and can the analysis can be referred to as univariate analysis.The bivariate analysis attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sales and spending can be considered as an example of bivariate analysis.The multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

Question 40: Explain Star Schema

.It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

Question 41: What is Cluster Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Question 42: What is Systematic Sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is equal probability method.

Question 43: What are Eigenvectors and Eigenvalues?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching.Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

Question 44: Can you cite some examples where a false positive is important than a false negative?

Let us first understand what false positives and false negatives are.False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.False Negatives are the

cases where you wrongly classify events as non-events, a.k.a Type II error.Example 1: In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn't have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.Example 2: Let's say an e-commerce company decided to give 1000Giftvouchertothecustomerswhomtheyassumetopurchaseatleast10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above 10,000.Nowtheissueisifwesendthe1000 gift vouchers to customers who have not actually purchased anything but are marked as having made $10,000 worth of purchase.

Question 45: Can you cite some examples where a false negative important than a false positive?

Example 1: Assume there is an airport 'A' which has received high-security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to a shortage of staff, they decide to scan passengers being predicted as risk positives by their predictive model. What will happen if a true threat customer is being flagged as non-threat by airport model?Example 2: What if Jury or judge decides to make a criminal go free?Example 3: What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

Question 46: Can you cite some examples where both false positive and false negatives are equally important?

In the Banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

Question 47: Can you explain the difference between a Validation Set and a Test Set?

A Validationset can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting of the model being built.On the other hand, a Test Set is used for testing or evaluating the performance of a trained machine learning model.In simple terms, the differences can be summarized as; training set is to fit the parameters i.e. weights and test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.

Question 48: Explain cross-validation.

Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will generalize to an independent dataset. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.