Basic Statistics Interview Questions

Ready to kickstart your Statistics career? This section is curated to help you understand the basics and has a list of basic statistics interview questions. Let's get started.

1. What is the Central Limit Theorem?

Central Limit Theorem is the cornerstone of statistics. It states that the distribution of a sample from a population comprising a large sample size will have its mean normally distributed. In other words, it will not have any effect on the original population distribution.

Central Limit Theorem is widely used in the calculation of confidence intervals and hypothesis testing. Here is an example – We want to calculate the average height of people in the world, and we take some samples from the general population, which serves as the data set. Since it is hard or impossible to obtain data regarding the height of every person in the world, we will simply calculate the mean of our sample.

By multiplying it several times, we will obtain the mean and their frequencies which we can plot on the graph and create a normal distribution. It will form a bell-shaped curve that will closely resemble the original data set.

2. What is the assumption of normality?

The assumption of normality dictates that the mean distribution across samples is normal. This is true across independent samples as well.

3. Describe Hypothesis Testing. How is the statistical significance of an insight assessed?

Hypothesis Testing in statistics is used to see if a certain experiment yields meaningful results. It essentially helps to assess the statistical significance of insight by determining the odds of the results occurring by chance. The first thing is to know the null hypothesis and then state it. Then the p-value is calculated, and if the null hypothesis is true, other values are also

determined. The alpha value denotes the significance and is adjusted accordingly.

If the p-value is less than alpha, the null hypothesis is rejected, but if it is greater than alpha, the null hypothesis is accepted. The rejection of the null hypothesis indicates that the results obtained are statistically significant.

4. What are observational and experimental data in statistics?

Observational data is derived from the observation of certain variables from observational studies. The variables are observed to determine any correlation between them.

Experimental data is derived from those experimental studies where certain variables are kept constant to determine any discrepancy or causality.

5. What is an outlier?

Outliers can be defined as the data points within a data set that varies largely in comparison to other observations. Depending on its cause, an outlier can decrease the accuracy as well as the efficiency of a model. Therefore, it is crucial to remove them from the data set.

6. How to screen for outliers in a data set?

There are many ways to screen and identify potential outliers in a data set. Two key methods are described below –

- Standard deviation/z-score – Z-score or standard score can be obtained in a normal distribution by calculating the size of one standard deviation and multiplying it by 3. The data points outside the range are then identified. The Z-score is measured from the mean. If the z-score is positive, it means the data point is above average.

If the z-score is negative, the data point is below average.

If the z-score is close to zero, the data point is close to average.

If the z-score is above or below 3, it is an outlier and the data point is considered unusual.

The formula for calculating a z-score is –

`z= data point-mean/standard deviation OR` $z=x-\mu/\sigma$

- Interquartile range (IQR) – IQR, also called midspread, is a method to identify outliers and can be described as the range of values that occur throughout the length of the middle of 50% of a data set. It is simply the difference between two extreme data points within the observation.

IQR=Q3 – Q1

Other methods to screen outliers include Isolation Forests, Robust Random Cut Forests, and DBScan clustering.

7. What is the meaning of an inlier?

An Inliner is a data point within a data set that lies at the same level as the others. It is usually an error and is removed to improve the model accuracy. Unlike outliers, inlier is hard to find and often requires external data for accurate identification.

8. What is the meaning of six sigma in statistics?

Six sigma in statistics is a quality control method to produce an error or defect-free data set. Standard deviation is known as Sigma or σ. The more the standard deviation, the less likely that process performs with accuracy and causes a defect. If a process outcome is 99.99966% error-free, it is considered six sigma. A six sigma model works better than 1σ, 2σ, 3σ, 4σ, 5σ processes and is reliable enough to produce defect-free work.

9. What is the meaning of KPI in statistics?

KPI is an acronym for a key performance indicator. It can be defined as a quantifiable measure to understand whether the goal is being achieved or not. KPI is a reliable metric to measure the performance level of an

organization or individual with respect to the objectives. An example of KPI in an organization is the expense ratio.

10. What is the Pareto principle?

Also known as the 80/20 rule, the Pareto principle states that 80% of the effects or results in an experiment are obtained from 20% of the causes. A simple example is – 20% of sales come from 80% of customers.

11. What is the Law of Large Numbers in statistics?

According to the law of large numbers, an increase in the number of trials in an experiment will result in a positive and proportional increase in the results coming closer to the expected value. As an example, let us check the probability of rolling a six-sided dice three times. The expected value obtained is far from the average value. And if we roll a dice a large number of times, we will obtain the average result closer to the expected value (which is 3.5 in this case).

12. What are some of the properties of a normal distribution?

Also known as Gaussian distribution, Normal distribution refers to the data which is symmetric to the mean, and data far from the mean is less frequent in occurrence. It appears as a bell-shaped curve in graphical form, which is symmetrical along the axes.

The properties of a normal distribution are –

- Symmetrical – The shape changes with that of parameter values
- Unimodal – Has only one mode.
- Mean – the measure of central tendency
- Central tendency – the mean, median, and mode lie at the centre, which means that they are all equal, and the curve is perfectly symmetrical at the midpoint.

13. How would you describe a ‘p-value’?

P-value in statistics is calculated during hypothesis testing, and it is a number that indicates the likelihood of data occurring by a random

chance. If a p-value is 0.5 and is less than alpha, we can conclude that there is a probability of 5% that the experiment results occurred by chance, or you can say, 5% of the time, we can observe these results by chance.

14. How can you calculate the p-value using MS Excel?

The formula used in MS Excel to calculate p-value is –

```
=tdist(x,deg_freedom,tails)
```

The p-value is expressed in decimals in Excel. Here are the steps to calculate it –

- Find the Data tab
- On the Analysis tab, click on the data analysis icon
- Select Descriptive Statistics and then click OK
- Select the relevant column
- Input the confidence level and other variables

15. What are the types of biases that you can encounter while sampling?

Sampling bias occurs when you lack the fair representation of data samples during an investigation or a survey. The six main types of biases that one can encounter while sampling are –

- Undercoverage bias
- Observer Bias
- Survivorship bias
- Self-Selection/Voluntary Response Bias
- Recall Bias
- Exclusion Bias

Intermediate Statistics Interview Questions

Planning to switch to a career where you need Statistics? This section will help you prepare well for the upcoming interview. It has a compiled list of intermediate statistics interview questions that are commonly asked during the interview process.

16. What is cherry-picking, P-hacking, and significance chasing?

Cherry-picking can be defined as the practice in statistics where only that information is selected which supports a certain claim and ignores any other claim that refutes the desired conclusion.

P-hacking refers to a technique in which data collection or analysis is manipulated until significant patterns can be found who have no underlying effect whatsoever.

Significance chasing is also known by the names of Data Dredging, Data Fishing, or Data Snooping. It refers to the reporting of insignificant results as if they are almost significant.

17. What is the difference between type I vs type II errors?

A type 1 error occurs when the null hypothesis is rejected even if it is true. It is also known as false positive.

A type 2 error occurs when the null hypothesis fails to get rejected, even if it is false. It is also known as a false negative.

18. What is a statistical interaction?

A statistical interaction refers to the phenomenon which occurs when the influence of an input variable impacts the output variable. A real-life example includes the interaction of adding sugar to the stirring of tea. Neither of the two variables has an impact on sweetness, but it is the combination of these two variables that do.

19. Give an example of a data set with a non-Gaussian distribution?

A non-Gaussian distribution is a common occurrence in many processes in statistics. This happens when the data naturally follows a non-normal distribution with data clumped on one side or the other on a graph. For example, the growth of bacteria follows a non-Gaussian or exponential distribution naturally and Weibull distribution.

20. What is the Binomial Distribution Formula?

The binomial distribution formula is:

```
b(x; n, P) = nCx * Px * (1 – P)n – x
```

Where:

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

21. What are the criteria that Binomial distributions must meet?

Here are the three main criteria that Binomial distributions must meet –

- The number of observation trials must be fixed. It means that one can only find the probability of something when done only a certain number of times.
- Each trial needs to be independent. It means that none of the trials should impact the probability of other trials.
- The probability of success remains the same across all trials.

22. What is linear regression?

In statistics, linear regression is an approach that models the relationship between one or more explanatory variables and one outcome variable. For example, linear regression can be used to quantify or model the relationship between various predictor variables such as age, gender, genetics, and diet on height, outcome variables.

23. What are the assumptions required for linear regression?

Four major assumptions for linear regression are as under –

- There's a linear relationship between the predictor (independent) variables and the outcome (dependent) variable. It means that the relationship between X and the mean of Y is linear.
- The errors are normally distributed with no correlation between them. This process is known as Autocorrelation.
- There is an absence of correlation between predictor variables. This phenomenon is called multicollinearity.
- The variation in the outcome or response variable is the same for all values of independent or predictor variables. This phenomenon of assumption of equal variance is known as homoscedasticity.

24. What are some of the low and high-bias Machine Learning algorithms?

Some of the widely used low and high-bias Machine Learning algorithms are –

Low bias -Decision trees, Support Vector Machines, k-Nearest Neighbors, etc.

High bias -Linear Regression, Logistic Regression, Linear Discriminant Analysis, etc.

Check out the free course on Statistical Methods For Decision Making.

25. When should you use a t-test vs a z-test?

The z-test is used for hypothesis testing in statistics with a normal distribution. It is used to determine population variance in the case where a sample is large.

The t-test is used with a t-distribution and used to determine population variance when you have a small sample size.

In case the sample size is large or n>30, a z-test is used. T-tests are helpful when the sample size is small or n<30.

26. What is the equation for confidence intervals for means vs for proportions?

To calculate the confidence intervals for mean, we use the following equation –

For n > 30

Use the Z table for the standard normal distribution.

For n<30

Use the t table with df=n-1

Confidence Interval for the Population Proportion –

27. What is the empirical rule?

In statistics, the empirical rule states that every piece of data in a normal distribution lies within three standard deviations of the mean. It is also known as the 68–95–99.7 rule. According to the empirical rule, the percentage of values that lie in a normal distribution follow the 68%, 95%, and 99.7% rule. In other words, 68% of values will fall within one standard deviation of the mean, 95% will fall within two standard deviations, and 99.75 will fall within three standard deviations of the mean.

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

Confidence tests and hypothesis tests both form the foundation of statistics.

The confidence interval holds importance in research to offer a strong base for research estimations, especially in medical research. The confidence interval provides a range of values that helps in capturing the unknown parameter.

Hypothesis testing is used to test an experiment or observation and determine if the results did not occur purely by chance or luck using the below formula where 'p' is some parameter.

Confidence and hypothesis testing are inferential techniques used to either estimate a parameter or test the validity of a hypothesis using a sample of data from that data set. While confidence interval provides a range of values for an accurate estimation of the precision of that parameter, hypothesis testing tells us how confident we are inaccurately drawing conclusions about a parameter from a sample. Both can be used to infer population parameters in tandem.

In case we include 0 in the confidence interval, it indicates that the sample and population have no difference. If we get a p-value that is higher than alpha from hypothesis testing, it means that we will fail to reject the bull hypothesis.

29. What general conditions must be satisfied for the central limit theorem to hold?

Here are the conditions that must be satisfied for the central limit theorem to hold –

- The data must follow the randomization condition which means that it must be sampled randomly.
- The Independence Assumptions dictate that the sample values must be independent of each other.
- Sample sizes must be large. They must be equal to or greater than 30 to be able to hold CLT. Large sample size is required to hold the accuracy of CLT to be true.

30. What is Random Sampling? Give some examples of some random sampling techniques.

Random sampling is a sampling method in which each sample has an equal probability of being chosen as a sample. It is also known as probability sampling.

Let us check four main types of random sampling techniques –

- Simple Random Sampling technique – In this technique, a sample is chosen randomly using randomly generated numbers. A sampling frame with the list of members of a population is required, which is denoted by 'n'. Using Excel, one can randomly generate a number for each element that is required.
- Systematic Random Sampling technique -This technique is very common and easy to use in statistics. In this technique, every k'th element is sampled. For instance, one element is taken from the sample and then the next while skipping the pre-defined amount or 'n'.

In a sampling frame, divide the size of the frame N by the sample size (n) to get 'k', the index number. Then pick every k'th element to create your sample.

- Cluster Random Sampling technique -In this technique, the population is divided into clusters or groups in such a way that each cluster represents the population. After that, you can randomly select clusters to sample.
- Stratified Random Sampling technique – In this technique, the population is divided into groups that have similar characteristics. Then a random sample can be taken from each group to ensure that different segments are represented equally within a population.

31. What is the difference between population and sample in inferential statistics?

A population in inferential statistics refers to the entire group we take samples from and are used to draw conclusions. A sample, on the other hand, is a specific group we take data from and this data is used to calculate the statistics. Sample size is always less than that of the population.

32. What are descriptive statistics?

Descriptive statistics are used to summarize the basic characteristics of a data set in a study or experiment. It has three main types –

- Distribution – refers to the frequencies of responses.

- Central Tendency – gives a measure or the average of each response.
- Variability – shows the dispersion of a data set.

33. What are quantitative data and qualitative data?

Qualitative data is used to describe the characteristics of data and is also known as Categorical data. For example, how many types. Quantitative data is a measure of numerical values or counts. For example, how much or how often. It is also known as Numeric data.

34. How to calculate range and interquartile range?

The range is the difference between the highest and the lowest values whereas the Interquartile range is the difference between upper and lower medians.

```
Range (X) = Max(X) – Min(X)
```

```
IQR = Q3 – Q1
```

Here, Q3 is the third quartile (75 percentile)

Here, Q1 is the first quartile (25 percentile)

35. What is the meaning of standard deviation?

Standard deviation gives the measure of the variation of dispersion of values in a data set. It represents the differences of each observation or data point from the mean.

```
(σ) = √(∑(x-μ)2 / n)
```

Where the variance is the square of standard deviation.

36. What is the relationship between mean and median in normal distribution?

In a normal distribution, the mean and the median are equal.

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

In the left-skewed distribution, the left tail is longer than the right side.

Mean < median < mode

In the right-skewed distribution, the right tail is longer. It is also known as positive-skew distribution.

Mode < median < mean

38. How to convert normal distribution to standard normal distribution?

Any point (x) from the normal distribution can be converted into standard normal distribution (Z) using this formula –

$$Z(standardized) = (x-\mu) / \sigma$$

Here, Z for any particular x value indicates how many standard deviations x is away from the mean of all values of x.

39. What can you do with an outlier?

Outliers affect A/B testing and they can be either removed or kept according to what situation demands or the data set requirements.

Here are some ways to deal with outliers in data –

- Filter out outliers especially when we have loads of data.
- If a data point is wrong, it is best to remove the outliers.
- Alternatively, two options can be provided – one with outliers and one without.
- During post-test analysis, outliers can be removed or modified. The best way to modify them is to trim the data set.

- If there are a lot of outliers and results are critical, then it is best to change the value of the outliers to other variables. They can be changed to a value that is representative of the data set.
- When outliers have meaning, they can be considered, especially in the case of mild outliers.

40. How to detect outliers?

The best way to detect outliers is through graphical means. Apart from that, outliers can also be detected through the use of statistical methods using tools such as Excel, Python, SAS, among others. The most popular graphical ways to detect outliers include box plot and scatter plot.

41. Why do we need sample statistics?

Sampling in statistics is done when population parameters are not known, especially when the population size is too large.

42. What is the relationship between standard error and margin of error?

Margin of error = Critical value X Standard deviation for the population

and

Margin of error = Critical value X Standard error of the sample.

The margin of error will increase with the standard error.

43. What is the proportion of confidence intervals that will not contain the population parameter?

Alpha is the probability in a confidence interval that will not contain the population parameter.

$$\alpha = 1 - CL$$

Alpha is usually expressed as a proportion. For instance, if the confidence level is 95%, then alpha would be equal to 1-0.95 or 0.05.

44. What is skewness?

Skewness provides the measure of the symmetry of a distribution. If a distribution is not normal or asymmetrical, it is skewed. A distribution can exhibit positive skewness or negative skewness if the tail on the right is longer and the tail on the left side is longer, respectively.

45. What is the meaning of covariance?

In statistics, covariance is a measure of association between two random variables from their respective means in a cycle.

46. What is a confounding variable?

A confounding variable in statistics is an ‘extra’ or ‘third’ variable that is associated with both the dependent variable and the independent variable, and it can give a wrong estimate that provides useless results.

For example, if we are studying the effect of weight gain, then lack of workout will be the independent variable, and weight gain will be the dependent variable. In this case, the amount of food consumption can be the confounding variable as it will mask or distort the effect of other variables in the study. The effect of weather can be another confounding variable that may later the experiment design.

47. What does it mean if a model is heteroscedastic?

A model is said to be heteroscedastic when the variation in errors comes out to be inconsistent. It often occurs in two forms – conditional and unconditional.

48. What is selection bias and why is it important?

Selection bias is a term in statistics used to denote the situation when selected individuals or a group within a study differ in a manner from the population of interest that they give systematic error in the outcome.

Typically, selection bias can be identified using bivariate tests apart from using other methods of multiple regression such as logistic regression.

It is crucial to understand and identify selection bias to avoid skewing results in a study. Selection bias can lead to false insights about a particular population group in a study.

Different types of selection bias include –

- Sampling bias – It is often caused by non-random sampling. The best way to overcome this is by drawing from a sample that is not self-selecting.
- Participant attrition – The dropout rate of participants from a study constitutes participant attrition. It can be avoided by following up with the participants who dropped off to determine if the attrition is due to the presence of a common factor between participants or something else.
- Exposure – It occurs due to the incorrect assessment or the lack of internal validity between exposure and effect in a population.
- Data – It includes dredging of data and cherry-picking and occurs when a large number of variables are present in the data causing even bogus results to appear significant.
- Time-interval – It is a sampling error that occurs when observations are selected from a certain time period only. For example, analyzing sales during the Christmas season.
- Observer selection- It is a kind of discrepancy or detection bias that occurs during the observation of a process and dictates that for the data to be observable, it must be compatible with the life that observes it.

49. What does autocorrelation mean?

Autocorrelation is a representation of the degree of correlation between the two variables in a given time series. It means that the data is correlated in a way that future outcomes are linked to past outcomes. Autocorrelation makes a model less accurate because even errors follow a sequential pattern.

50. What does Design of Experiments mean?

The Design of Experiments or DOE is a systematic method that explains the relationship between the factors affecting a process and its output. It is used to infer and predict an outcome by changing the input variables.

51. What is Bessel's correction?

Bessel's correction advocates the use of n-1 instead of n in the formula of standard deviation. It helps to increase the accuracy of results while analyzing a sample of data to derive more general conclusions.

52. What types of variables are used for Pearson's correlation coefficient?

Variables (both the dependent and independent variables) used for Pearson's correlation coefficient must be quantitative. It will only test for the linear relationship between two variables.

53. What is the use of Hash tables in statistics?

In statistics, hash tables are used to store key values or pairs in a structured way. It uses a hash function to compute an index into an array of slots in which the desired elements can be searched.

54. Does symmetric distribution need to be unimodal?

Symmetrical distribution does not necessarily need to be unimodal, they can be skewed or asymmetric. They can be bimodal with two peaks or multimodal with multiple peaks.

55. What is the benefit of using box plots?

Boxplot is a visually effective representation of two or more data sets and facilitates quick comparison between a group of histograms.

56. What is the meaning of TF/IDF vectorization?

TF/IDF is an acronym for Term Frequency – Inverse Document Frequency and is a numerical measure widely used in statistics in summarization. It

reflects the importance of a word or term in a document. The document is called a collection or corpus.

57. What is the meaning of sensitivity in statistics?

Sensitivity refers to the accuracy of a classifier in a test. It can be calculated using the formula –

Sensitivity = Predicted True Events/Total number of Events

58. What is the difference between the first quartile, the second quartile, and the third quartile?

The first quartile is denoted by Q1 and it is the median of the lower half of the data set.

The second quartile is denoted by Q2 and is the median of the data set.

The third quartile is denoted by Q3 and is the median of the upper half of the data set.

About 25% of the data set lies above Q3, 75% lies below Q3 and 50% lies below Q2. The Q1, Q2, and Q3 are the 25^{th}, $50^{th,}$ and 75^{th} percentile respectively.

59. What is kurtosis?

Kurtosis is a measure of the degree of the extreme values present in one tail of distribution or the peaks of frequency distribution as compared to the others. The standard normal distribution has a kurtosis of 3 whereas the values of symmetry and kurtosis between -2 and +2 are considered normal and acceptable. The data sets with a high level of kurtosis imply that there is a presence of outliers. One needs to add data or remove outliers to overcome this problem. Data sets with low kurtosis levels have light tails and lack outliers.

60. What is a bell-curve distribution?

A bell-curve distribution is represented by the shape of a bell and indicates normal distribution. It occurs naturally in many situations especially while analyzing financial data. The top of the curve shows the mode, mean and median of the data and is perfectly symmetrical. The key characteristics of a bell-shaped curve are –

- The empirical rule says that approximately 68% of data lies within one standard deviation of the mean in either of the directions.
- Around 95% of data falls within two standard deviations and
- Around 99.7% of data fall within three standard deviations in either direction.

Statistics FAQs

How do I prepare for a statistics interview?
To prepare for a statistics interview, you can read this blog on the top commonly asked interview questions. These questions will help you brush up your skills and ace your upcoming interview.

What are the most important topics in statistics?
Estimation: bias, maximum likelihood, method of moments, Rao-Blackwell theorem, fisher information. Central limit theorem, hypothesis testing, likelihood ratio tests, law of large numbers – These are some of the most important topics in statistics.

What are basics of statistics?
A collection of methods to display, analyze, and draw conclusions from data. Statistics can be of two types, descriptive statistics and inferential statistics.

What are the 7 steps in hypothesis testing?
1. State the null hypothesis
2. State the alternate hypothesis
3. Which test and test statistic to be performed
4. Collect Data
5. Calculate the test statistic
6. Construct Acceptance / Rejection regions
7. Based on steps 5 and 6, draw a conclusion about H0

Basic Interview Questions

1. How is the statistical significance of an insight assessed?

Hypothesis testing is used to find out the statistical significance of the insight. To elaborate, the null hypothesis and the alternate hypothesis are stated, and the p-value is calculated.

After calculating the p-value, the null hypothesis is assumed true, and the values are determined. To fine-tune the result, the alpha value, which denotes the significance, is tweaked. If the p-value turns out to be less than the alpha, then the null hypothesis is rejected. This ensures that the result obtained is statistically significant.

2. Where are long-tailed distributions used?

A long-tailed distribution is a type of distribution where the tail drops off gradually toward the end of the curve.

The Pareto principle and the product sales distribution are good examples to denote the use of long-tailed distributions. Also, it is widely used in classification and regression problems.

3. What is the central limit theorem?

The central limit theorem states that the normal distribution is arrived at when the sample size varies without having an effect on the shape of the population distribution.

This central limit theorem is the key because it is widely used in performing hypothesis testing and also to calculate the confidence intervals accurately.

Learn more about Quantitative Methods with our blog!

4. What is observational and experimental data in Statistics?

Observational data correlates to the data that is obtained from observational studies, where variables are observed to see if there is any correlation between them.

Experimental data is derived from experimental studies, where certain variables are held constant to see if any discrepancy is raised in the working.

Check out our blog on Statistics for Data Science!

5. What is meant by mean imputation for missing data? Why is it bad?

Mean imputation is a rarely used practice where null values in a dataset are replaced directly with the corresponding mean of the data.

It is considered a bad practice as it completely removes the accountability for feature correlation. This also means that the data will have low variance and increased bias, adding to the dip in the accuracy of the model, alongside narrower confidence intervals.

6. What is an outlier? How can outliers be determined in a dataset?

Outliers are data points that vary in a large way when compared to other observations in the dataset. Depending on the learning process, an outlier can worsen the accuracy of a model and decrease its efficiency sharply.

Outliers are determined by using two methods:

- Standard deviation/z-score
- Interquartile range (IQR)

7. How is missing data handled in statistics?

There are many ways to handle missing data in Statistics:

- Prediction of the missing values
- Assignment of individual (unique) values
- Deletion of rows, which have the missing data
- Mean imputation or median imputation
- Using random forests, which support the missing values

8. What is exploratory data analysis?

Exploratory data analysis is the process of performing investigations on data to understand the data better.

In this, initial investigations are done to determine patterns, spot abnormalities, test hypotheses, and also check if the assumptions are right.

9. What is the meaning of selection bias?

Selection bias is a phenomenon that involves the selection of individual or grouped data in a way that is not considered to be random. Randomization plays a key role in performing analysis and understanding model functionality better.

If correct randomization is not achieved, then the resulting sample will not accurately represent the population.

10. What are the types of selection bias in statistics?

There are many types of selection bias as shown below:

- Observer selection
- Attrition
- Protopathic bias
- Time intervals

- Sampling bias

11. What is the meaning of an inlier?

An inlier is a data point that lies at the same level as the rest of the dataset. Finding an inlier in the dataset is difficult when compared to an outlier as it requires external data to do so. Inliers, similar to outliers reduce model accuracy. Hence, even they are removed when they're found in the data. This is done mainly to maintain model accuracy at all times.

12. What is the probability of getting a sum of 5 or 8 when 2 dice are rolled once?

When 2 dice are rolled,

Total outcomes = 36 (i.e. 6*6)

Possible outcomes of getting 5 = 4

Possible outcomes of getting a sum 8 = 5

Total = 9

Probability =9/36 = 1/4 = 0.25

13. State the case where the median is a better measure when compared to the mean.

In the case where there are a lot of outliers that can positively or negatively skew data, the median is preferred as it provides an accurate measure in this case of determination.

14. Can you give an example of root cause analysis?

Root cause analysis, as the name suggests, is a method used to solve problems by first identifying the root cause of the problem.

Example: If the higher crime rate in a city is directly associated with the higher sales in a red-colored shirt, it means that they are having a positive correlation. However, this does not mean that one causes the other.

Causation can always be tested using A/B testing or hypothesis testing.

15. What is the meaning of six sigma in statistics?

Six sigma is a quality assurance methodology used widely in statistics to provide ways to improve processes and functionality when working with data.

A process is considered as six sigma when 99.99966% of the outcomes of the model are considered to be defect-free.

16. What is DOE?

DOE is an acronym for the Design of Experiments in statistics. It is considered as the design of a task that describes the information and the change of the same based on the changes to the independent input variables.

17. What is the meaning of KPI in statistics?

KPI stands for Key Performance Analysis in statistics. It is used as a reliable metric to measure the success of a company with respect to its achieving the required business objectives.

There are many good examples of KPIs:

- Profit margin percentage
- Operating profit margin
- Expense ratio

18. What type of data does not have a log-normal distribution or a Gaussian distribution?

Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well.

Example: Duration of a phone car, time until the next earthquake, etc.

19. What is the Pareto principle?

The Pareto principle is also called the 80/20 rule, which means that 80 percent of the results are obtained from 20 percent of the causes in an experiment.

A simple example of the Pareto principle is the observation that 80 percent of peas come from 20 percent of pea plants on a farm.

20. What is the meaning of the five-number summary in Statistics?

The five-number summary is a measure of five entities that cover the entire range of data as shown below:

- Low extreme (Min)
- First quartile (Q1)
- Median
- Upper quartile (Q3)
- High extreme (Max)

21. What are population and sample in Inferential Statistics, and how are they different?

A population is a large volume of observations (data). The sample is a small portion of that population. Because of the large volume of data in the population,

it raises the computational cost. The availability of all data points in the population is also an issue.

In short:

- We calculate the statistics using the sample.
- Using these sample statistics, we make conclusions about the population.

22. What are quantitative data and qualitative data?

- Quantitative data is also known as numeric data.
- Qualitative data is also known as categorical data.

23. What is Mean?

Mean is the average of a collection of values. We can calculate the mean by dividing the sum of all observations by the number of observations.

24. What is the meaning of standard deviation?

Standard deviation represents the magnitude of how far the data points are from the mean. A low value of standard deviation is an indication of the data being close to the mean, and a high value indicates that the data is spread to extreme ends, far away from the mean.

25. What is a bell-curve distribution?

A normal distribution can be called a bell-curve distribution. It gets its name from the bell curve shape that we get when we visualize the distribution.

26. What is skewness?

Skewness measures the lack of symmetry in a data distribution. It indicates that there are significant differences between the mean, the mode, and the median of data. Skewed data cannot be used to create a normal distribution.

27. What is kurtosis?

Kurtosis is used to describe the extreme values present in one tail of distribution versus the other. It is actually the measure of outliers present in the distribution. A high value of kurtosis represents large amounts of outliers being present in data. To overcome this, we have to either add more data into the dataset or remove the outliers.

28. What is correlation?

Correlation is used to test relationships between quantitative variables and categorical variables. Unlike covariance, correlation tells us how strong the relationship is between two variables. The value of correlation between two variables ranges from -1 to +1.

The -1 value represents a high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will drastically decrease. Similarly, +1 means a positive correlation, and here, an increase in one variable will lead to an increase in the other. Whereas, 0 means there is no correlation.

If two variables are strongly correlated, then they may have a negative impact on the statistical model, and one of them must be dropped.

Next up on this top Statistics Interview Questions and Answers blog, let us take a look at the intermediate set of questions.

Intermediate Interview Questions

29. What are left-skewed and right-skewed distributions?

A left-skewed distribution is one where the left tail is longer than that of the right tail. Here, it is important to note that the mean < median < mode.

Similarly, a right-skewed distribution is one where the right tail is longer than the left one. But, here mean > median > mode.

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

Descriptive Statistics: Descriptive statistics is used to summarize a sample set of data like the standard deviation or the mean.

Inferential statistics: Inferential statistics is used to draw conclusions from the test data that are subjected to random variations.

31. What are the types of sampling in Statistics?

There are four main types of data sampling as shown below:

- **Simple random**: Pure random division
- **Cluster**: Population divided into clusters
- **Stratified**: Data divided into unique groups
- **Systematical**: Picks up every 'n' member in the data

32. What is the meaning of covariance?

Covariance is the measure of indication when two items vary together in a cycle. The systematic relation is determined between a pair of random variables to see if the change in one will affect the other variable in the pair or not.

33. Imagine that Jeremy took part in an examination. The test is having 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?

To determine the solution to the problem, the following formula is used:

```
X = μ + Zσ


Here:

μ: Mean

σ: Standard deviation

X: Value to be calculated


Therefore, X = 160 + (15*1.2) = 173.8 (Approximated to 174)
```

If you are looking forward to becoming an expert in Statistics and Data Analytics, make sure to check out Intellipaat's **online Data Analyst Course** program.

34. If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?

If the given distribution is a right-skewed distribution, then the mean should be greater than 20, while the mode remains to be less than 20.

35. What is Bessel's correction?

Bessel's correction is a factor that is used to estimate a populations' standard deviation from its sample. It causes the standard deviation to be less biased, thereby, providing more accurate results.

36. The standard normal curve has a total area to be under one, and it is symmetric around zero. True or False?

True, a normal curve will have the area under unity and the symmetry around zero in any distribution. Here, all of the measures of central tendencies are equal to zero due to the symmetric nature of the standard normal curve.

37. 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?

First, correlation does not imply causation here. Correlation is only used to measure the relationship, which is linear between rest and productive work. If both vary rapidly, then it means that there is a high amount of correlation between them.

38. What is the relationship between the confidence level and the significance level in statistics?

The significance level is the probability of obtaining a result that is extremely different from the condition where the null hypothesis is true. While the confidence level is used as a range of similar values in a population.

Both significance and confidence level are related by the following formula:

Significance level = 1 − Confidence level

39. A regression analysis between apples (y) and oranges (x) resulted in the following least-squares line: y = 100 + 2x. What is the implication if oranges are increased by 1?

If the oranges are increased by one, there will be an increase of 2 apples since the equation is:

```
y = 100 + 2x.
```

40. What types of variables are used for Pearson's correlation coefficient?

Variables to be used for the Pearson's correlation coefficient must be either in a ratio or in an interval.

Note that there can exist a condition when one variable is a ratio, while the other is an interval score.

41. In a scatter diagram, what is the line that is drawn above or below the regression line called?

The line that is drawn above or below the regression line in a scatter diagram is called the residual or also the prediction error.

42. What are the examples of symmetric distribution?

Symmetric distribution means that the data on the left side of the median is the same as the one present on the right side of the median.

There are many examples of symmetric distribution, but the following three are the most widely used ones:

- Uniform distribution
- Binomial distribution
- Normal distribution

43. Where is inferential statistics used?

Inferential statistics is used for several purposes, such as research, in which we wish to draw conclusions about a population using some sample data. This is performed in a variety of fields, ranging from government operations to quality control and quality assurance teams in multinational corporations.

44. What is the relationship between mean and median in a normal distribution?

In a normal distribution, the mean is equal to the median. To know if the distribution of a dataset is normal, we can just check the dataset's mean and median.

45. What is the difference between the Ist quartile, the IInd quartile, and the IIIrd quartile?

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.

That is,

- **The lower quartile (Q1)** is the 25th percentile.
- **The middle quartile (Q2)**, also called the median, is the 50th percentile.
- **The upper quartile (Q3)** is the 75th percentile.

46. How do the standard error and the margin of error relate?

The standard error and the margin of error are quite closely related to each other. In fact, the margin of error is calculated using the standard error. As the standard error increases, the margin of error also increases.

47. What is one sample t-test?

This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population's mean.

48. What is an alternative hypothesis?

The alternative hypothesis (denoted by H1) is the statement that must be true if the null hypothesis is false. That is, it is a statement used to contradict the null hypothesis. It is the opposing point of view that gets proven right when the null hypothesis is proven wrong.

Check out this Data Science Certification course to become a certified Data Scientist.

49. Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?

Given that it is a left-skewed distribution, the mean will be less than the median, i.e., less than 60, and the mode will be greater than 60.

50. What are the types of biases that we encounter while sampling?

Sampling biases are errors that occur when taking a small sample of data from a large population as the representation in statistical analysis. There are three types of biases:

- The selection bias
- The survivorship bias
- The undercoverage bias

Next up on this top Statistics Interview Questions and answers blog, let us take a look at the advanced set of questions.

Advanced Interview Questions

51. What are the scenarios where outliers are kept in the data?

There are not many scenarios where outliers are kept in the data, but there are some important situations when they are kept. They are kept in the data for analysis if:

- Results are critical
- Outliers add meaning to the data
- The data is highly skewed

52. Briefly explain the procedure to measure the length of all sharks in the world.

Following steps can be used to determine the length of sharks:

- Define the confidence level (usually around 95%)
- Use sample sharks to measure
- Calculate the mean and standard deviation of the lengths
- Determine t-statistics values
- Determine the confidence interval in which the mean length lies

53. How does the width of the confidence interval change with length?

The width of the confidence interval is used to determine the decision-making steps. As the confidence level increases, the width also increases.

The following also apply:

- Wide confidence interval: Useless information
- Narrow confidence interval: High-risk factor

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

Degrees of freedom or DF is used to define the number of options at hand when performing an analysis. It is mostly used with t-distribution and not with the z-distribution.

If there is an increase in DF, the t-distribution will reach closer to the normal distribution. If DF > 30, this means that the t-distribution at hand is having all of the characteristics of a normal distribution.

55. How can you calculate the p-value using MS Excel?

Following steps are performed to calculate the p-value easily:

- Find the Data tab above
- Click on Data Analysis
- Select Descriptive Statistics
- Select the corresponding column
- Input the confidence level

56. What is the law of large numbers in statistics?

The law of large numbers in statistics is a theory that states that the increase in the number of trials performed will cause a positive proportional increase in the average of the results becoming the expected value.

Example: The probability of flipping a fair coin and landing heads is closer to 0.5 when it is flipped 100,000 times when compared to 100 flips.

57. What are some of the properties of a normal distribution?

A normal distribution, regardless of its size, will have a bell-shaped curve that is symmetric along the axes.

Following are some of the important properties:

- Unimodal: It has only one mode.
- Symmetrical: Left and right halves of the curve are mirrored.
- Central tendency: The mean, median, and mode are at the midpoint.

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

The probability of not seeing a supercar in 20 minutes is:

```
= 1 – P(Seeing one supercar)
= 1 – 0.3
= 0.7
```

Probability of not seeing any supercar in the period of 60 minutes is:

```
= (0.7) ^ 3 = 0.343
```

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

```
= 1 – P(Not seeing any supercar)
= 1 – 0.343 = 0.657
```

59. What is the meaning of sensitivity in statistics?

Sensitivity, as the name suggests, is used to determine the accuracy of a classifier (logistic, random forest, etc.):

The simple formula to calculate sensitivity is:

Sensitivity = Predicted True Events/Total number of Events

60. What are the types of biases that you can encounter while sampling?

There are three types of biases:

- Selection bias
- Survivorship bias
- Under coverage bias

61. What is the meaning of TF/IDF vectorization?

TF-IDF is an acronym for Term Frequency – Inverse Document Frequency. It is used as a numerical measure to denote the importance of a word in a document. This document is usually called the collection or the corpus.

The TF-IDF value is directly proportional to the number of times a word is repeated in a document. TF-IDF is vital in the field of Natural Language Processing (NLP) as it is mostly used in the domain of text mining and information retrieval.

62. What are some of the low and high-bias Machine Learning algorithms?

There are many low and high-bias Machine Learning algorithms, and the following are some of the widely used ones:

- **Low bias**: **SVM**, decision trees, KNN algorithm, etc.
- **High bias**: Linear and logistic regression

Check out this Machine Learning Training in Noida and master ML Skills.

63. What is the use of Hash tables in statistics?

Hash tables are the data structures that are used to denote the representation of key-value pairs in a structured way. The hashing function is used by a hash table to compute an index that contains all of the details regarding the keys that are mapped to their associated values.

64. What are some of the techniques to reduce underfitting and overfitting during model training?

Underfitting refers to a situation where data has high bias and low variance, while overfitting is the situation where there are high variance and low bias.

Following are some of the techniques to reduce underfitting and overfitting:

For reducing underfitting:

- Increase model complexity
- Increase the number of features
- Remove noise from the data
- Increase the number of training epochs

For reducing overfitting:

- Increase training data
- Stop early while training
- Lasso regularization
- Use random dropouts

65. Can you give an example to denote the working of the central limit theorem?

Let's consider the population of men who have normally distributed weights, with a mean of 60 kg and a standard deviation of 10 kg, and the probability needs to be found out.

If one single man is selected, the weight is greater than 65 kg, but if 40 men are selected, then the mean weight is far more than 65 kg.

The solution to this can be as shown below:

```
Z = (x – μ) / ? = (65 – 60) / 10 = 0.5


For a normal distribution P(Z > 0.5) = 0.409
Z = (65 – 60) / 5 = 1
P(Z > 1) = 0.090
```

66. How do you stay up-to-date with the new and upcoming concepts in statistics?

This is a commonly asked question in a statistics interview. Here, the interviewer is trying to assess your interest and ability to find out and learn new things efficiently. Do talk about how you plan to learn new concepts and make sure to elaborate on how you practically implemented them while learning.

If you are looking forward to learning and mastering all of the Data Analytics and Data Science concepts and earn a certification in the same, do take a look at Intellipaat's latest **Data Science with R Certification** offerings.

67. What is the benefit of using box plots?

Box plots allow us to provide a graphical representation of the 5-number summary and can also be used to compare groups of histograms.

Check out this Python Data Science Course to get an in-depth understanding of Data Science and Python.

68. Does a symmetric distribution need to be unimodal?

A symmetric distribution does not need to be unimodal (having only one mode or one value that occurs most frequently). It can be bi-modal (having two values that have the highest frequencies) or multi-modal (having multiple or more than two values that have the highest frequencies).

69. What is the impact of outliers in statistics?

Outliers in statistics have a very negative impact as they skew the result of any statistical query. For example, if we want to calculate the mean of a dataset that contains outliers, then the mean calculated will be different from the actual mean (i.e., the mean we will get once we remove the outliers).

70. When creating a statistical model, how do we detect overfitting?

Overfitting can be detected by cross-validation. In cross-validation, we divide the available data into multiple parts and iterate on the entire dataset. In each iteration, one part is used for testing, and others are used for training. This way, the entire dataset will be used for training and testing purposes, and we can detect if the data is being overfitted.

71. What is a survivorship bias?

The survivorship bias is the flaw of the sample selection that occurs when a dataset only considers the 'surviving' or existing observations and fails to consider those observations that have already ceased to exist.

72. What is an undercoverage bias?

The undercoverage bias is a bias that occurs when some members of the population are inadequately represented in the sample.

74. What is the relationship between standard deviation and standard variance?

Standard deviation is the square root of standard variance. Basically, standard deviation takes a look at how the data is spread out from the mean. On the other hand, standard variance is used to describe how much the data varies from the mean of the entire dataset.