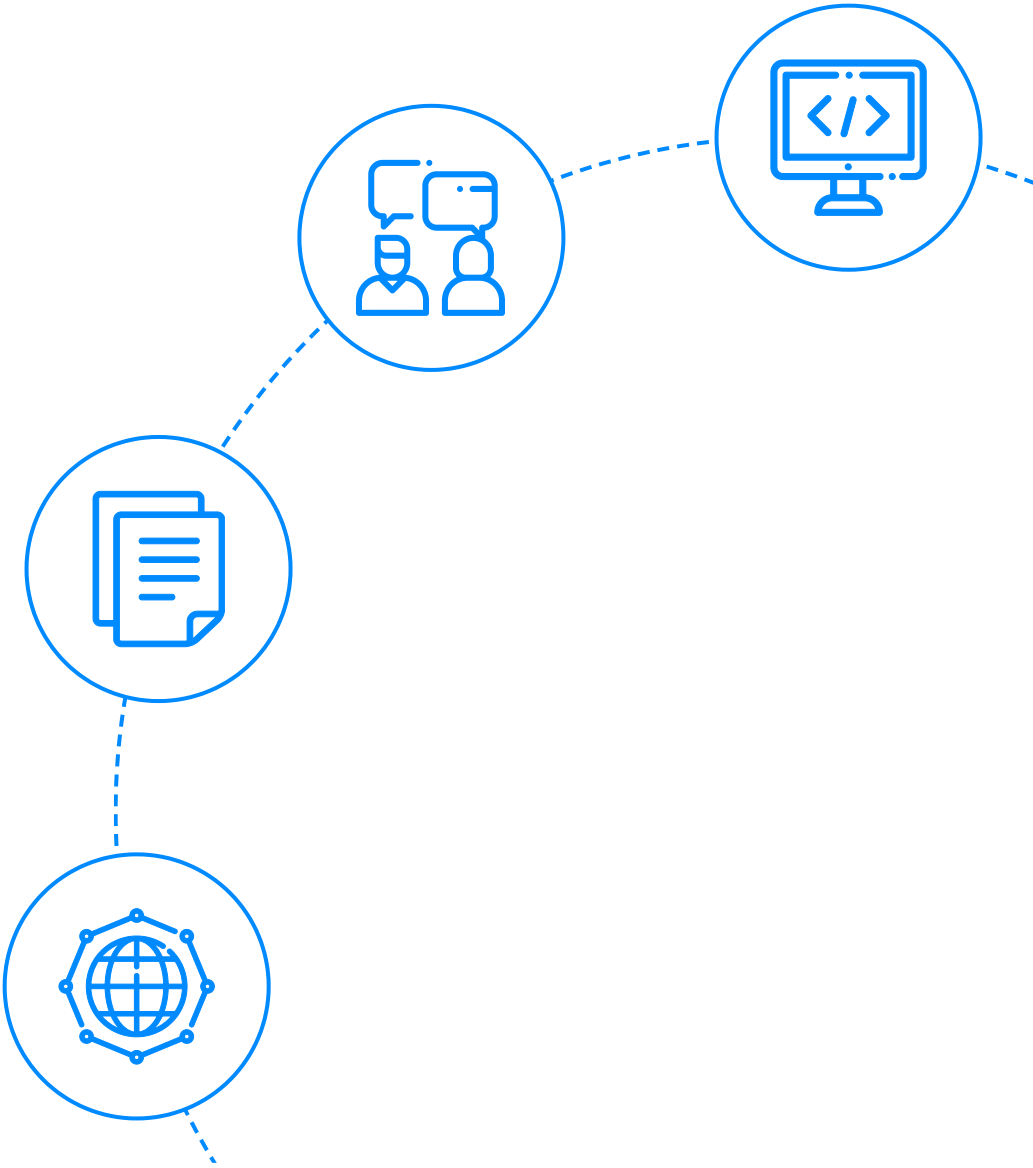
InterviewBit

Statistics Interview Questions



To view the live version of the page, click here.

Contents

Basic Statistics Interview Questions for Freshers

Statistics Interview Questions for Experienced

21. What is sampling?
22. How do you determine the statistical significance of an insight?
23. How does the central limit theorem work?
24. How do you define exploratory data analysis?
25. What is the definition of selection bias?
26. What does it mean by inlier?
27. What are the applications of long-tailed distributions?
28. In what situation would the median be a more suitable measure compared to the mean?
29. How might root cause analysis be applied to a real-life situation?
30. How does the Design of Experiments in statistics work?
31. What does standard deviation mean?
32. What are the characteristics of a bell-curve distribution?
33. What is your definition of skewness?
34. How do you define kurtosis?
35. What is the definition of correlation?
36. Left-skewed and right-skewed distributions exist, what are they?
37. How does the term covariance relate to understanding?
38. How do you define Bessel's correction?
39. What are inferential statistics used for?
40. How are mean and median related in a normal distribution?

Statistics Interview Questions for Experienced

41. What is the relationship between standard error and the margin of error?
42. What does a degree of freedom (DF) represent in statistics?
43. How do you explain the law of large numbers in statistics?
44. How does TF/IDF vectorization relate to meaning?
45. What is the purpose of Hash tables in statistics?
46. Symmetric distributions need to be unimodal, does it?
47. Is there any significance to outliers in statistics?
48. What is the meaning of Central Tendency?
49. How do you define Normal Distribution?
50. How do you define empirical rule?

Let's get Started

Statistics have a significant impact on today's computing and data handling, and many companies invest billions of dollars into it. This is a very interesting area of study, and many organisations use Analytics.

Let’s see what is Statistics in simple words, Statistics is a fundamental area of data science that provides a strong foundation for learning data science and processing large quantities of data. Every field of research uses statistics, which is composed of collecting, analysing, interpreting, and presenting a huge amount of numerical data. **Data science** relies heavily on statistics. It ensures that the data is understood.

Because of this, we've put together the most common **Statistics Interview Questions and Answers**. Every interview is different, so it is not feasible to give you one formula. We have provided answers to the most common statistics questions to assist you in your job interview. We will cover Hypothesis testing, central limit theorem, Six sigma, KPI, error, p-values, biased etc.

The following answers can help you prepare for the Statistics Interviews.

Basic Statistics Interview Questions for Freshers

1. What is Statistics?

Statistics is the discipline that studies and develops techniques for gathering, processing, analyzing, interpreting, and communicating statistical information (using information gathered from research).

2. What is the central limit theorem?

The central limit theorem is the foundation of statistics. It states that if a sample is drawn from a population with large sample size, the distribution of the sample's mean will be distributed normally. In other words, the original population distribution will not be affected.

The central limit theorem is extremely useful in estimating confidence intervals and testing hypotheses. For instance, let's say I want to estimate the worldwide average height. I would take a sample of people from the general population and calculate the mean. Because it is difficult or impossible to collect data on every person's height, the mean of my sample will serve as my estimate.

To create a normal curve, we can plot the mean value and the frequency on a graph and then multiply them several times. The resulting curve will be similar to the original data set, but it will be slightly shifted to the left.

3. What is a hypothesis test? How is the statistical significance of an insight determined?

The statistical significance of an experiment's insights can be assessed using hypothesis testing. Hypothesis testing examines the probability of a given experiment's results occurring by chance. The null hypothesis is defined first, and then p-values are computed. If the null hypothesis is true, other values are determined as well. As its name suggests, the alpha value indicates the degree of significance.

In a two-tailed test, the p-value is less than alpha if the null hypothesis is rejected but is greater than alpha if the null hypothesis is accepted. In a one-tailed test, the p-value is less than alpha if the null hypothesis is accepted but is greater than alpha if the null hypothesis is rejected. The rejection of the null hypothesis indicates that the results obtained are statistically significant.

4. Why are statistical data referred to as observational and experimental?

Correlations between variables can be discovered through the collection of observational data.

To determine the cause or effect of a particular variable, experimental data is collected from those experiments where it is kept constant.

5. What is the definition of an inlier?

An error instance is usually identified as an Inliner within a data set. It is usually a lower-level data point and should therefore be removed. Finding inliers is usually difficult and requires outside data to identify.

6. What is the Six sigma in statistic?

In quality control, an error-free data set is generated using six sigma statistics. σ is known as standard deviation. The lower the standard deviation, the less likely that a process performs accurately and commits errors. If a process delivers 99.99966% error-free results, it is said to be six sigma. A six sigma model is one that outperforms 1σ, 2σ, 3σ, 4σ, and 5σ processes and is sufficiently reliable to deliver defect-free work.

7. What does KPI stand for in statistics?

- A KPI is a quantifiable measure to evaluate whether the objectives are being met or not.
- It is a reliable metric to measure the performance level of an organisation or individual.
- An example of a KPI in an organisation such as the expense ratio.
- In terms of performance, KPIs are an effective way of measuring whether an organisation or individual is meeting expectations.

8. Why is the Pareto principle famous?

The Pareto principle states that 80% of the effects or results in an experiment come from 20% of the causes. The Pareto principle is often applied to business to explain that 80% of the profits or results come from 20% of the efforts. To illustrate, 80% of customers buy 20% of the items.

9. What are the characteristics of large numbers in statistics?

When the number of trials in an experiment increases, the results will approach the expected value in a desirable proportion because of the law of large numbers. To determine the probability of rolling a six-sided die three times, we can use this example. The outcome is far from the expected value, and if we roll the die a large number of times, we will more likely obtain our desired result closer to the expected value (3.5 in this instance).

10. What do you think of the phrase 'p-value'?

It is a number that helps determine the probability of a random occurrence when evaluating a hypothesis. In statistics, the p-value indicates how likely it is that a particular dataset occurred by chance. If the p-value is less than alpha, we can conclude that there is a probability of 5% that the experiment results occurred by chance or 5% of the time, we would see these results.

11. What are cherry-picking, P-hacking, and significance chasing?

- Cherry-picking is the act of exclusively taking the bits of information that support a particular conclusion and ignoring all the bits of information that contradict it.
- P-hacking, also known as data collection or analysis manipulation, is a technique that produces significant patterns even though they have no underlying effect.
- Reporting insignificant results as if they are almost significant, is known as Significance Chasing. Data Dredging, Data Fishing, and Data Snooping are all names for this behaviour.

12. What is the difference between an error of type I and an error of type II?

- When the null hypothesis is rejected even though it is correct, a type 1 error occurs. False positives are also known as type 1 errors.
- When the null hypothesis is not rejected despite being incorrect, a type 2 error occurs. This is also known as a false negative.

13. How does one define statistical interaction?

- When an input variable influences an output variable, a statistical interaction occurs.
- In real life, for example, the interaction of adding sugar to the stirring of tea is an example of statistical interaction. Neither of the variables has an impact on sweetness, but the two variables combine to produce sweetness.

14. What are some examples of data sets with non-Gaussian distributions?

When data follows a non-normal distribution, it is frequently non-Gaussian. A non-Gaussian distribution is often seen in many statistics processes. This occurs when data is naturally clustered on one side or the other on a graph. For instance, bacterial growth follows an exponential or non-Gaussian distribution, which is non-normal.

15. How does linear regression work?

When utilised in statistics, linear regression is a technique that models the relationship between one or more predictor variables and one outcome variable. For example, linear regression may be used to study the connection between various predictors, such as age, gender, heredity, diet, and height.

16. What are the necessary conditions for a Binomial Distribution?

The three most important characteristics of a Binomial Distribution are listed below.

1. The number of observations must be prearranged. In other words, one can only determine the probability of an event happening a specific number of times if a fixed number of trials are performed.
2. It is important that each trial is independent of the others. This means that the probability of each subsequent trial should not be affected by previous trials.
3. The chance of getting the job remains the same no matter how many times you try.

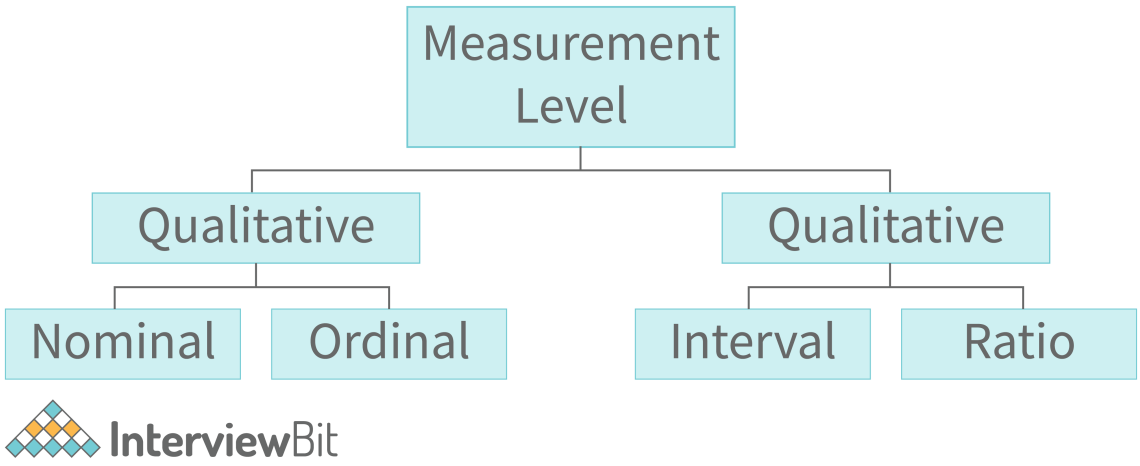
17. What is the difference between a sample and a population?

The subset of the population from which numbers are obtained is known as the sample. The numbers obtained from the population are known as parameters, while the numbers obtained from the sample are known as statistics. It is through sample data that conclusions may be made about the population.

Population	Sample
A parameter is an observable quality that can be measured.	Statistics is an observable quality that can be measured.
Every element of the population is a unique individual.	A subset of the population is used to explore some aspects of the population.
An opinion report is a true representation of what happened.	The reported values have a confidence level and an error margin.
All members of a group are included in the list.	A particular portion of the population is represented by that subset.

18. What are the different kinds of variables or levels of measurement?

A variable can be categorized as one of four types: Ordinal, Interval, Ratio, or Nominal. Scale and Continuous are sometimes used to describe Interval and Ratio levels of measurement, respectively.



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

Descriptive	Inferential
Describe the data in terms of its key characteristics.	To conclude the population, it is used.
Data can be organised, analysed, and presented in a meaningful way thanks to charts.	The purpose of data analysis is to compare data and make predictions through hypotheses.
Using charts, tables, and graphs to present information.	Probability was responsible for achieving this goal.

20. What are the differences between Data Science and Statistics?

Data Science, which is a scientific discipline that employs data, includes interdisciplinary methods, algorithms, and even the procedure for extracting data knowledge. The data can be either coded or uncoerced. Data mining and data science are similar because both provide abstract data from a large amount of data. Data science now includes computer science, mathematical statistics, and computer science and behavioural applications. Data science, which integrates data analysis, understanding, organization, and communication together, produces insights and knowledge from a large amount of data by combining statistical analysis, visualization, and applied mathematical economics. The collection, analysis, interpretation, organization, and presentation of data are the main components of data science.

The points listed below provide information on the differences between data science and statistics:

On Comparison	Data Science	Statistics
Definition	• A branch of scientific techniques that is related to disciplines other than science. • Data mining involves processes, algorithms, and systems, too. • To obtain information from data, use structured or unstructured data.	• A set of operations for dealing with data is provided. • Part of the mathematics branch. • Design experiments can be accomplished by providing methods. • The plan collects data, analyzes it, and forms an image to represent it for additional assessments.

Statistics Interview Questions for Experienced

21. What is sampling?

Selecting an unbiased or random subset of individual observations in a population is regarded as part of the statistical practice of sampling. In order to obtain some understanding of the population, sampling is used.

22. How do you determine the statistical significance of an insight?

- The p-value is used to determine whether the null hypothesis is true or false. To put it another way, the null hypothesis states that there is no difference between the conditions, and the alternate hypothesis states that there is a difference. The p-value is then calculated.
- Once the p-value has been calculated, the null hypothesis is accepted and the sample values are determined. The alpha value, which indicates the significance of the result, is adjusted to fine-tune the result. If the p-value is lower than the alpha, the null hypothesis is rejected and the result is statistically significant.

23. How does the central limit theorem work?

A stable distribution is one whose parameters change only slightly when the sample size changes. The central limit theorem says that a normal distribution is a result when the sample size is unchanged and the population shape doesn't change.

The central limit theorem is crucial because it provides us with the correct formula for calculating confidence intervals. It is also used to test hypotheses correctly.

24. How do you define exploratory data analysis?

The goal of an exploratory data analysis is to better understand data by conducting investigations on it. During this stage, patterns are detected, hypotheses are tested, anomalies are spotted, and the foundation for the research is established.

25. What is the definition of selection bias?

The process of selecting individual or group data in a way that is not random is known as selection bias. Randomization is crucial in evaluating model functionality and performing analysis. Therefore, if incorrect randomization is not avoided, the obtained sample will not accurately represent the population.

26. What does it mean by inlier?

Finding an inlier in a dataset is more challenging than finding an outlier. Because finding an inlier requires external data, model accuracy is often maintained. The reduction in model accuracy caused by the presence of inliers is the same as that caused by outliers. Even when inliers are detected in the data, they are usually removed to maintain model accuracy.

27. What are the applications of long-tailed distributions?

The part of the curve that extends to the end is known as a long tail. It gradually gets smaller towards the end of the curve.

The long-tailed distribution is used to demonstrate the Pareto principle and the product sales distribution in these examples. It is also utilised in classification and regression problems.

28. In what situation would the median be a more suitable measure compared to the mean?

In situations where outliers can affect data in either a positive or negative manner, the median is preferable due to its ability to accurately gauge this.

29. How might root cause analysis be applied to a real-life situation?

- The technique of identifying the source of a problem by identifying the root cause is known as root cause analysis.
- **Examples:** A positive correlation between the higher crime rate in a city and the higher sales of red shirts can be inferred from the above sentence. However, this does not mean that one causes the other.
- Correlational and experimental approaches can always be used to test causation.

30. How does the Design of Experiments in statistics work?

The Design of Experiments in Statistics is an experimental design that defines an inquiry task that specifies how variable changes when another variable change. It is also known as the **Design of Experiments**.

31. What does standard deviation mean?

When a set of data points is near the mean, a low value of standard deviation indicates that the points are close to the mean, and a high value indicates that the points are far away from the mean. On the other hand, when the data points are far apart from each other, a high standard deviation indicates that the points are far away from the mean, and a low standard deviation indicates that the points are close to the mean.

32. What are the characteristics of a bell-curve distribution?

The characteristic bell curve shape of a normal distribution is what gives it its name. We can perceive the bell curve as we look at the distribution.

33. What is your definition of skewness?

Skewed data distribution has a non-symmetrical pattern relative to the mean, the mode, and the median. The skewness of data indicates that there are significant differences between the mean, the mode, and the median. Data that is skewed cannot be used to create a normal distribution.

34. How do you define kurtosis?

Outliers are detected in a data distribution using kurtosis. It measures the extent to which the tail values diverge from the central portion of the distribution. The higher the kurtosis, the higher the number of outliers in the data. To reduce their effect, we may either include more data or eliminate the outliers.

35. What is the definition of correlation?

- The degree to which variables correlate is tested by covariance and correlation. In contrast to covariance, correlation indicates how closely linked two variables are. Values for correlation range from -1 to +1, with -1 indicating a strong negative correlation and +1 indicating a strong positive correlation.
- A high negative correlation, where if one variable increases, the other variable will decrease drastically, is represented by the -1 value. A positive correlation, where an increase in one variable will cause an increase in the other, is represented by the +1 value. There is no correlation between 0 and +1 variables, whereas 0 and -1 variables have a negative correlation.
- If the statistical model is affected negatively by two variables that are strongly correlated, then one of them must be removed.

36. Left-skewed and right-skewed distributions exist, what are they?

- The left tail is longer than the right tail in a left-skewed distribution. It is critical to note here that mean, median, and mode are inverses of one another.
- In contrast to a left-skewed distribution, in which the left tail is longer than the right one, a right-skewed distribution is one where the right tail is longer than the left one. Here, the mean > the median > the mode.

37. How does the term covariance relate to understanding?

When two items are associated in a random process, covariance is the measure of how closely they fluctuate together. Is there a connection between one of the variables in a random pair and the other variable? If there is, then the systematic connection is determined.

38. How do you define Bessel's correction?

Bessel's correction corrects the flaw in using a sample to estimate a population standard deviation. It lowers the bias in the estimated standard deviation, resulting in more accurate measurements.

39. What are inferential statistics used for?

In inferential statistics, we use some sample data to draw conclusions about a population. From government operations to quality control and quality assurance teams in multinational corporations, inferential statistics are used in a variety of fields.

40. How are mean and median related in a normal distribution?

The mean and the median of a dataset are in agreement if the dataset's distribution is normal. We can immediately tell if a dataset's distribution is normal if we simply check its mean and median.

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

The margin of error is proportionally influenced by the standard error. In other words, the margin of error is computed using standard error. As standard error increases, the margin of error also rises.

42. What does a degree of freedom (DF) represent in statistics?

The t-distribution is used to calculate degrees of freedom and not the z-distribution. When speaking about degrees of freedom, we are referring to the number of options at our disposal when conducting an analysis.

The t-distribution will shift closer to a normal distribution as DF increases. If DF is greater than 30, this means that the t-distribution at hand has all of the characteristics of a normal distribution.

43. How do you explain the law of large numbers in statistics?

Inference from statistical data can be said to follow the law of large numbers, which purports that, as the number of trials increases, the average result will increase in proportion to it. The percentage of heads obtained by repeatedly flipping a fair coin is lower the more times it is flipped, 100,000 times in this example.

44. How does TF/IDF vectorization relate to meaning?

A numerical value representing the importance of a word in a document is referred to as TF-IDF. It is measured using the Term Frequency – Inverse Document Frequency

The phrase frequency-inverse document frequency value is directly proportional to the number of times a word appears in a document. Text mining and information retrieval are mainly dependent on phrase frequency-inverse document frequency values.

45. What is the purpose of Hash tables in statistics?

When key-value pairs are stored in a hash table, the information regarding keys and associated values are stored in a hierarchical fashion using hash tables. The hashing function is used to provide an index that contains all of the information regarding keys and their associated values.

46. Symmetric distributions need to be unimodal, does it?

Bi- or multi-modal symmetric distributions do not have to have only one mode or value that occurs most frequently, nor do they have to be unimodal (having only one mode or value that occurs most frequently).

47. Is there any significance to outliers in statistics?

Outliers have a significant detrimental impact on the calculation of any statistical query result. For example, if we seek to compute the mean of a dataset that contains outliers, we will get a different result than the actual mean (i.e., the mean we would get after removing the outliers).

48. What is the meaning of Central Tendency?

The central tendency measures (signifies) the central position within the dataset by referencing a single value. The three most common central tendency measures are the mean, the median, and the mode.

1. Mean: The Arithmetic Mean is the sum of all values divided by the number of values. In cases where n values are provided (x1, x2, x3, …, xn), the following formula can be used.

$$\text{Mean } (\bar{x}) = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

2. Median: The median is the value located in the middle when the data is ordered (e.g. ordered in ascending or descending order). If n values are provided (such as x1, x2, x3,……xn), then the median is x ⁀ = { x1, x2, x3,……xn } .

- If n is odd, the case is I.

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

- If n is even, the case is II.

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2} + 1\right)^{th}}{2}$$

i.e mean of two middle values

3. Mode: In the dataset, there may be more than one value that is the mode. Therefore, the mode is the most frequent value.

49. How do you define Normal Distribution?

The mean of a Normal Distribution is located symmetrically about the distribution's centre. It is also known as a Gaussian Distribution. A symmetric curve resembles a bell, with the most frequent data occurring at the centre (see the figure).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

where

σ : Standard Deviation

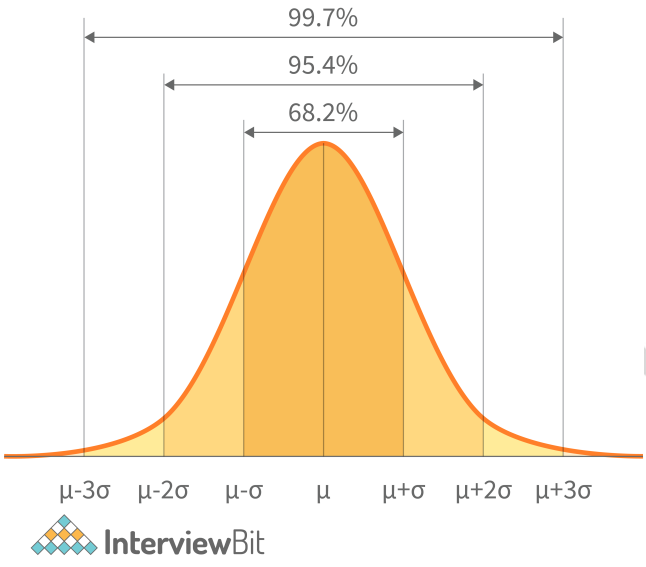
μ : Mean

x : random variable

50. How do you define empirical rule?

The 68 – 95 – 99.7 rule or the Three Sigma Rule refers to the proposition that on a Normal Distribution,

- There will be 68% within one Standard Error of the Mean of the data.
- There will be 95% of the data within two Standard deviations of the mean.
- There is around a 97% chance that the data will be within three standard deviations of the mean.



There's a lot of work involved in preparing for a data science interview, no matter how much experience you've gained or how impressive your statistics degree is. An interview may surprise you no matter how much work you've done or how many statistics courses you've taken, so be sure to keep those questions in mind when you discuss.

The Statistics Interview Questions and Answers listed here provide a fundamental understanding of Statistics from the ground up to the most advanced concepts. They enable students and professionals to get a comprehensive view of the field.

That concludes our interview questions on statistics! Hopefully, this has refreshed your knowledge on the subject.

Useful Resources

- https://www.interviewbit.com/technical-interview-questions/
- https://www.interviewbit.com/coding-interview-questions/
- https://www.interviewbit.com/mock-interview/
- https://www.interviewbit.com/blog/

Links to More Interview Questions

C Interview Questions	Php Interview Questions	C Sharp Interview Questions
Web Api Interview Questions	Hibernate Interview Questions	Node Js Interview Questions
Cpp Interview Questions	Oops Interview Questions	Devops Interview Questions
Machine Learning Interview Questions	Docker Interview Questions	Mysql Interview Questions
Css Interview Questions	Laravel Interview Questions	Asp Net Interview Questions
Django Interview Questions	Dot Net Interview Questions	Kubernetes Interview Questions
Operating System Interview Questions	React Native Interview Questions	Aws Interview Questions
Git Interview Questions	Java 8 Interview Questions	Mongodb Interview Questions
Dbms Interview Questions	Spring Boot Interview Questions	Power Bi Interview Questions
Pl Sql Interview Questions	Tableau Interview Questions	Linux Interview Questions
Ansible Interview Questions	Java Interview Questions	Jenkins Interview Questions