INTERVIEW

Top 25 Statistics Interview Questions for Data Science [2025]

By Meghana D

Apr 14, 2025 | 7 Min Read | 4864 Views
(Last Updated)

Are you about to sit for a data science interview but confused about the statistics bit because it seems like too much math? Because same! I remember when I was preparing for my data science interviews and just did not even want to look at the stats bit.

But statistics is the backbone of data science interviews, yet many of us prioritize coding skills alone. Your statistical expertise will determine your success in data science interviews, especially when you explain descriptive statistics that summarize data or inferential statistics that make predictions about populations.

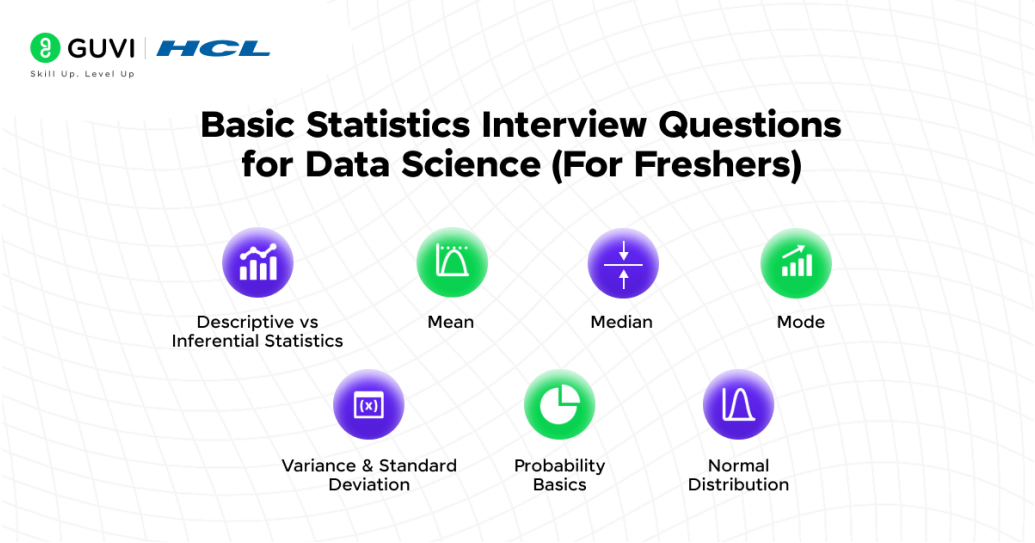
Hence, so that you don't have to be as confused as I was, I have drafted this guide that features 25 must-know statistics interview

questions for data science with answers. The questions range from simple concepts for freshers to advanced topics for experienced professionals. Each question includes a detailed yet concise answer to help you prepare well for your next data science interview. Let's begin.

Table of contents

Beginner-Level Statistics Interview Questions for Data Science (For Freshers)

Statistics is the foundation of data science. It supports analytical techniques and machine learning algorithms. Data science freshers need to learn fundamental statistical concepts to crack their interviews.



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

Answer:
A population refers to the entire set of individuals or data points that we want to study, while a sample is a subset of that population, chosen for analysis. Since analyzing an entire population is often impractical due to time and resource constraints, we use statistical sampling techniques to draw meaningful conclusions.

For example, if you want to analyze the average height of students in a country, measuring every student (population) would be difficult. Instead, you could take a sample from different schools and use statistical methods to estimate the average height of all students.

2. What is the difference between descriptive and inferential statistics?

Answer:

- **Descriptive statistics** summarize and present data in a meaningful way using measures such as mean, median, mode, variance, and standard deviation. It helps in understanding the basic characteristics of data without drawing conclusions beyond the dataset.

- **Inferential statistics** uses data from a sample to make predictions or generalizations about a larger population. It involves techniques like hypothesis testing, confidence intervals, and regression analysis.

For example, calculating the average salary of employees in a company using data from all employees is descriptive statistics, whereas using a sample of employees to predict salary trends for the whole industry is inferential statistics.

3. What are the different types of data in statistics?

Answer:
In statistics, data is categorized into four types:

1. **Nominal Data:** Categorical data without any order (e.g., colors, gender, city names).
2. **Ordinal Data:** Categorical data with a meaningful order but without precise differences (e.g., rankings like *poor, average, good*).
3. **Interval Data:** Numeric data with equal intervals but no true zero point (e.g., temperature in Celsius or Fahrenheit).
4. **Ratio Data:** Numeric data with a true zero point and equal intervals (e.g., height, weight, age, income).

Understanding these data types is crucial for selecting the appropriate statistical methods and visualizations.

4. What are mean, median, and mode? Why are they important?

Answer:
These are measures of central tendency used to summarize data:

- **Mean (Average):** The sum of all values divided by the number of observations. It is affected by extreme values (outliers).

- **Median:** The middle value when data is arranged in ascending order. It is useful when data has outliers since it is not influenced by extreme values.
- **Mode:** The most frequently occurring value in a dataset. It is useful for categorical data.

For example, in a dataset of ages {18, 20, 22, 22, 25, 30, 35}, the mean is 24, the median is 22, and the mode is 22.

5. What is standard deviation, and why is it important?

Answer:
Standard deviation (SD) measures the amount of variation or dispersion in a dataset. A low SD indicates that data points are close to the mean, while a high SD suggests that data points are spread out over a wide range.

For example, in two classes where students' test scores are:

- **Class A:** 80, 82, 83, 85, 87 (low variation, low SD)
- **Class B:** 60, 70, 80, 90, 100 (high variation, high SD)

The standard deviation helps you understand data consistency and reliability in data science.

6. What is probability, and what are its types?

Answer:
Probability is a measure of how likely an event is to occur, ranging from 0 (impossible event) to 1 (certain event).

There are three main types of probability:

1. **Classical Probability:** Based on known possible outcomes (e.g., rolling a fair die, the probability of getting a six is 1/6).
2. **Frequentist Probability:** Based on experimental outcomes (e.g., if a coin is flipped 100 times and lands on heads 60 times, the probability is 60/100).
3. **Bayesian Probability:** Updated as new evidence is introduced (used in machine learning and AI).

Understanding probability helps in decision-making, especially in predictive modeling.

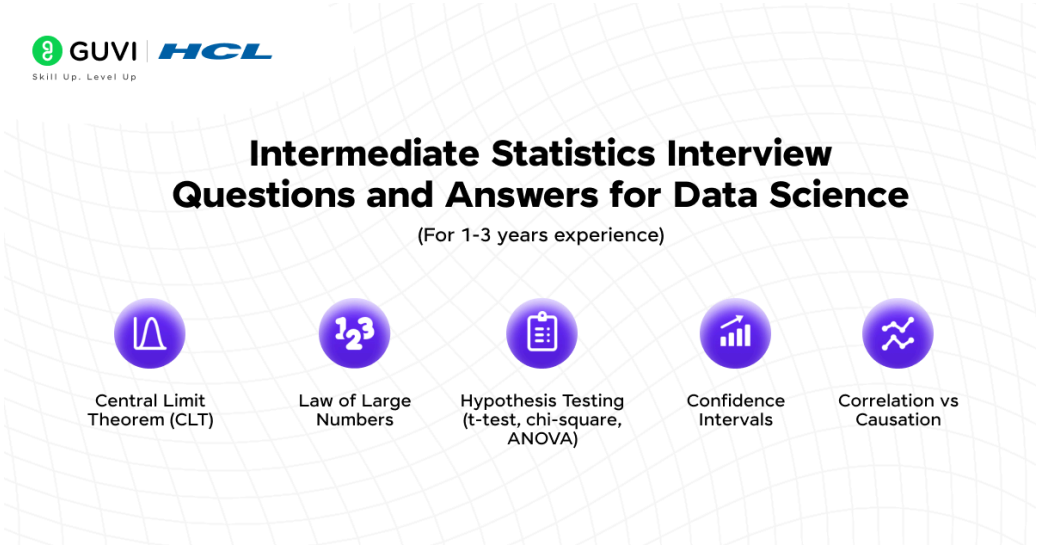
7. What is the Law of Large Numbers?

Answer:
The Law of Large Numbers (LLN) states that as the number of trials in an experiment increases, the observed results tend to get closer to the expected probability.

For example, if you flip a fair coin 10 times, you may not get exactly 5 heads and 5 tails, but if you flip it 10,000 times, the ratio of heads to tails will approach 50-50. This principle is crucial in statistical sampling and machine learning algorithms.

Intermediate-Level Statistics Interview Questions for Data Science (1-3 years experience)

Data scientists moving up from entry-level roles need to build a stronger statistical foundation. Interviewers expect candidates with 1-3 years of experience to show deep knowledge of statistics and its ground applications in business problems.



1. What is Hypothesis Testing, and Why is it Important?

Hypothesis testing is a statistical method used to make decisions based on sample data. It involves formulating two hypotheses:

- **Null Hypothesis (H_0):** Assumes there is no significant effect or difference.
- **Alternative Hypothesis (H_1):** Suggests that there is a significant effect or difference.

A test statistic is calculated, and a p-value is used to determine whether to reject H_0. If the p-value is less than the significance level (e.g., 0.05), we reject the null hypothesis. Hypothesis testing is critical in data science as it helps validate assumptions and make data-driven decisions.

2. What is the Difference Between Type I and Type II Errors?

- **Type I Error (False Positive):** Rejecting a true null hypothesis. Example: A medical test incorrectly detects a disease in a healthy person.
- **Type II Error (False Negative):** Failing to reject a false null hypothesis. Example: A test fails to detect a disease in a sick

person.

Minimizing Type I errors is important in high-risk fields like medicine, whereas reducing Type II errors is crucial in fraud detection to avoid missing real fraudulent activities.

3. What is a Confidence Interval?

A confidence interval (CI) represents a range within which a population parameter (like the mean) is expected to fall with a certain confidence level (e.g., 95% CI).

For example, if a 95% CI for the mean salary of data scientists is ₹10L – ₹15L, it means we are 95% confident that the actual mean salary falls within this range. Larger sample sizes lead to narrower confidence intervals, increasing the precision of our estimates.

4. When Should You Use a t-Test vs. a z-Test?

- **t-Test:** Used when the **sample size is small (n < 30)** and the population standard deviation is unknown.
- **z-Test:** Used when the **sample size is large (n ≥ 30)** and the population standard deviation is known.

For instance, if we are comparing the average test scores of two student groups (n = 20 each), a t-test is used. However, if the groups are large (n = 100+ each) with a known population variance, a z-test is preferred.

5. Explain Linear Regression and Its Assumptions

Linear regression is a method for modeling the relationship between a dependent variable (Y) and one or more independent variables (X) using a straight-line equation:

$Y=\beta 0+\beta 1X+\varepsilon$ $Y = \beta\_0 + \beta\_1X + \varepsilon$ $Y=\beta 0+\beta 1X+\varepsilon$

Where:

- $\beta 0$ β_0 $\beta 0$ is the intercept

- β1β_1β1 is the coefficient (slope)
- εεε is the error term

Key assumptions of linear regression:

1. **Linearity:** The Relationship between X and Y is linear.
2. **Independence:** Observations are independent of each other.
3. **Homoscedasticity:** Equal variance of residuals across all levels of X.
4. **Normality of Errors:** Residuals follow a normal distribution.

If these assumptions are violated, the model's predictions may be unreliable.

6. What is Multicollinearity, and How Do You Detect It?

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, leading to unreliable estimates of regression coefficients.

Detection Methods:

- **Variance Inflation Factor (VIF):** If VIF > 10, multicollinearity is high.
- **Correlation Matrix:** Variables with a correlation > 0.8 indicate potential multicollinearity.

Solution:

- Remove one of the correlated variables.
- Use **Principal Component Analysis (PCA)** to reduce dimensionality.

7. What is Overfitting in Machine Learning, and How Can You Prevent It?

Overfitting in machine learning happens when a model learns noise instead of the actual pattern, leading to excellent performance on training data but poor generalization to new data.

Ways to prevent overfitting:

- Use **cross-validation (e.g., k-fold CV)**.
- Apply **regularization techniques** like Lasso and Ridge Regression.
- Gather more data to reduce model variance.

For instance, in a spam detection model, an overfitted model may memorize specific words rather than recognize broader spam patterns.

8. How Do You Interpret an ROC Curve and AUC?

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (TPR) vs. False Positive Rate (FPR) at different threshold levels in a classification model.

- **AUC (Area Under the Curve):** Measures the classifier's ability to distinguish between classes.
- **AUC = 1.0:** Perfect classifier.
- **AUC = 0.5:** Random guessing (poor model).

For example, an AUC of 0.85 indicates a strong classifier with an 85% probability of distinguishing between positive and negative cases.

9. What is Bootstrapping in Statistics?

Bootstrapping is a resampling technique used to estimate population parameters by randomly sampling data with replacement. It helps in confidence interval estimation and model validation when the sample size is small.

For example, in a small dataset of 50 customer reviews, bootstrapping can generate multiple datasets by resampling to estimate the overall customer sentiment with greater accuracy.

10. What is the Central Limit Theorem (CLT), and why is it important?

Answer:
The Central Limit Theorem (CLT) states that, regardless of the original population distribution, the sampling distribution of the sample mean will tend to be normal (bell-shaped) if the sample size is sufficiently large (typically n > 30).

For example, if we take repeated random samples of the heights of people from different countries, the average heights of these samples will form a normal distribution. CLT is the foundation for hypothesis testing and confidence intervals.

11. What is correlation, and how is it different from causation?

Answer:
Correlation measures the strength and direction of a relationship between two variables. It is represented by the correlation coefficient (r), which ranges from -1 to +1:

- **r = +1:** Strong positive correlation (both variables increase together).
- **r = −1:** Strong negative correlation (one variable increases, the other decreases).
- **r = 0:** No correlation.

Causation, on the other hand, implies that one variable directly influences another. For example, ice cream sales and drowning incidents are correlated (both increase in summer), but eating ice cream does not cause drowning—temperature is the actual cause.

12. What is a p-value in hypothesis testing?

Answer:
A **p-value** helps determine the significance of statistical results. It represents the probability of observing the data if the null hypothesis (H_0) is true.

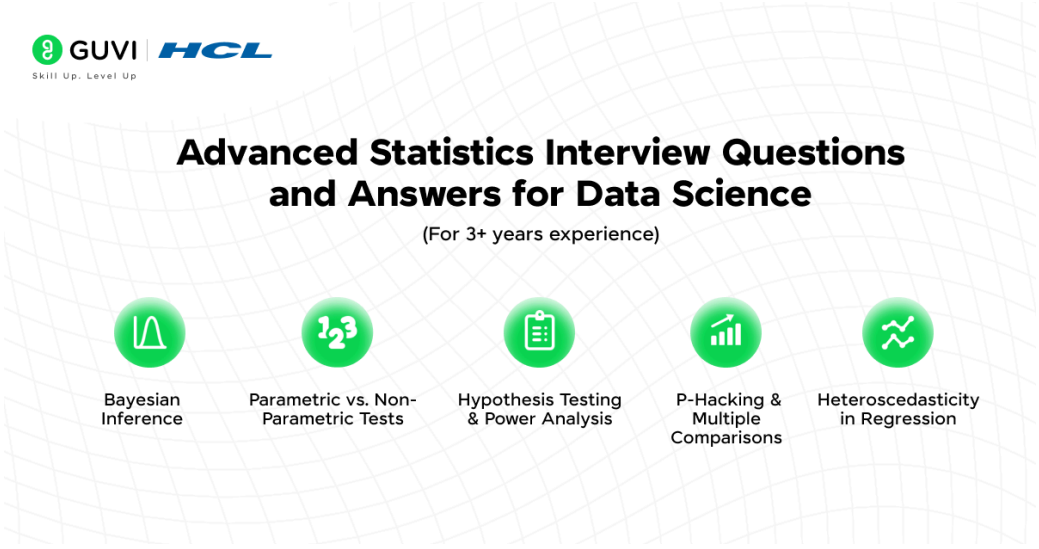
- A **low p-value (< 0.05)** suggests strong evidence **against** the null hypothesis (we reject H_0).
- A **high p-value (> 0.05)** indicates weak evidence against H_0 (we fail to reject it).

For example, in a drug effectiveness test, if the p-value = 0.02, it means there is a 2% probability that the observed results happened by chance, so we reject H_0 and conclude that the drug has a significant effect.

Looking to master data science and ace your statistics interviews? Enroll in GUVI's Data Science Course, designed to equip you with industry-relevant skills in Python, machine learning, statistics, and AI. With live mentorship, real-world projects, and placement support, this program helps you build a strong foundation for a successful data science career

Advanced-Level Statistics Interview Questions for Data Science (3+ years experience)

Data scientists with 3+ years of experience need statistical expertise that goes beyond simple methods. They should know complex methodologies to tackle sophisticated data challenges. Statistical theory questions in interviews test how well you can apply your knowledge to real-life scenarios.



1. Explain maximum likelihood estimation and its applications in data science

Answer:

Maximum likelihood estimation (MLE) helps estimate the parameters of statistical models by finding values that maximize the likelihood function. MLE serves as the foundation for many machine learning algorithms like logistic regression and neural networks.

The implementation of MLE boils down to finding parameter values that make your observed data most likely. Here's what you need to do:

1. Define a likelihood function based on your data distribution
2. Take the logarithm (to make calculations easier)
3. Find parameter values that maximize this function

MLE helps determine word probabilities in language models in natural language processing, to name just one example. In spite of that, MLE can be sensitive to outliers and might overfit when the data is limited.

2. What is the difference between Parametric and Non-Parametric Statistical Tests?

Answer:
Parametric tests assume that the data follows a known distribution (usually a normal distribution) and rely on parameters like mean and standard deviation. Examples include t-tests, ANOVA, and linear regression.

Non-parametric tests make fewer assumptions about data distribution and are useful when the data does not meet normality conditions. Examples include the Mann-Whitney U test, the Kruskal-Wallis test, and the Chi-square test.

Non-parametric tests are often used for ordinal data or when sample sizes are small and the distribution is unknown.

3. What is the Concept of Statistical Power in Hypothesis Testing?

Answer:
Statistical power is the probability that a test correctly rejects the null hypothesis when the alternative hypothesis is true (i.e., avoiding a Type II error). It is given by:

Power=1-β

where β is the probability of a Type II error (false negative).

Higher power means a greater likelihood of detecting an actual effect. Power depends on:

- **Sample size:** Larger samples increase power.
- **Effect size:** Larger effects are easier to detect.
- **Significance level (α):** Lowering α\alphaα reduces power.

To ensure a reliable test, a power of **at least 0.8 (80%)** is recommended.

4. What is P-Hacking and How to Avoid It?

Answer:

P-hacking refers to the misuse of statistical methods to obtain statistically significant ($p < 0.05$) results by selectively reporting data, running multiple tests, or adjusting analysis methods until a desired outcome is found.

To avoid P-hacking:

- Use pre-registered hypotheses to commit to an analysis plan before seeing the data.
- Correct for multiple comparisons using Bonferroni or Holm-Bonferroni corrections.
- Report effect sizes and confidence intervals instead of relying only on p-values.

- Use Bayesian statistics, which naturally incorporates prior evidence.

5. What is Heteroscedasticity in Regression Analysis?

Answer:
Heteroscedasticity occurs when the variance of errors in a regression model is not constant across all levels of an independent variable. This violates the assumption of homoscedasticity, which is required for valid inference in linear regression.

It can be detected using:

- **Residual plots** (increasing spread of residuals).
- **Breusch-Pagan test** or **White test**.

To fix heteroscedasticity:

- **Transform variables** (e.g., log transformation).
- **Use robust standard errors** or **weighted least squares (WLS)** regression.

6. How Do You Interpret the ROC Curve and AUC?

Answer:
The Receiver Operating Characteristic (ROC) Curve is a graphical plot of a classifier's True Positive Rate (TPR) vs. False Positive Rate (FPR) at different threshold levels.

The Area Under the Curve (AUC) represents the classifier's ability to distinguish between classes:

- **AUC = 1.0** → Perfect classifier.
- **AUC = 0.5** → Random guessing.
- **AUC < 0.5** → Worse than random.

A high AUC value means better model performance in classification problems.

Concluding Thoughts...

As we've established, statistics form a core pillar of data science that professionals must master throughout their careers. Experience level should guide your preparation approach.

New professionals should focus on core statistical concepts and their practical applications. Mid-level practitioners need deeper knowledge of hypothesis testing, regression analysis, and sampling methods. Senior data scientists must excel at advanced techniques like Bayesian modeling and causal inference. This is just a gist of what your learning journey should be like.

Your success in data science depends on mastering the skills at your respective level and staying current with statistical methods, along with applying them practically. I hope this article has aided your learning journey and if you have any doubts about the questions, do reach out to me in the comments section below.

FAQs

Q1. What are the key statistical concepts every data scientist should know?

Q2. How can I prepare for statistics questions in a data science interview?

Q3. What's the difference between descriptive and inferential statistics?

Q4. How does statistical knowledge apply to machine learning?

Success Stories

View All Stories

About the Author

Meghana D

I am a technical content writer with professional experience creating engaging and innovative content. My expertise includes writing about various technical topics to establish a strong brand presence online.

View all posts by Meghana D

Did you enjoy this article?

Name

Email

Rate this Article ☆ ☆ ☆ ☆ ☆

Enter Your Comment

Submit My Comment

Get In Touch For Details!

Name*

Email ID*

Phone Number*

+91

Education Qualification*

Choose qualification

Current Profile*

Select profile

Select your interested program*

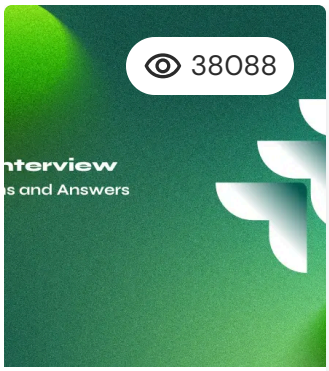
Choose Program

Get In Touch Request Information

Recommended Courses

Guvi Zen Class Guvi Courses

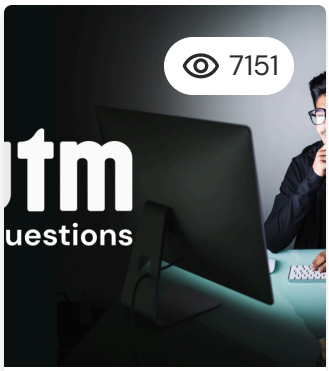
Similar Articles



Top 45 JIRA Interview Questions and Answers (2025)

By Lavish Jain

Top 30 Power BI Developer Interview Questions and Answers



10 Important Paytm Coding Interview Questions

By Archana

Blog Categories

Data Science

Artificial Intelligence and Machine Learning

Full Stack Development

DevOps

Digital Marketing

UI/UX Designing

VFX

Interview Questions

General Interview Questions

Java Interview Questions

Python Interview Questions

SQL Interview Questions

Selenium Interview Questions

Data Science Interview Questions

Salary blog

UI/UX Designer Salary

Data Scientist Salary

Full-Stack Developer Salary

Motion Graphics Designer Salary

Cloud Computing Engineer Salary

Digital Marketer Salary

About us

Our Story

Careers

Refund Policy

FAQs

Contact Us

Table of contents