KDnuggets

Machine Learning
MLOps
NLP
Programming
Python
SQL

Datasets
Events
Resources
Cheat Sheets
Recommendations
Tech Briefs

Advertise

JOIN NEWSLETTER

10 Statistics Questions to Ace Your Data Science Interview

Master statistics and land your first data science job.

By **Natassha Selvaraj**, KDnuggets Technical Content Specialist At-Large on August 8, 2024 in **Data Science**

Image by Author

I am a data scientist with a background in computer science.

I’m familiar with data structures, object oriented programming, and database management since I was taught these concepts for 3 years in university.

Blog
Top Posts
About

Topics
AI
Career Advice
Computer Vision
Data Engineering
Data Science
Language Models

However, when entering the field of data science, I noticed a significant skill gap.

I didn’t have the math or statistics background required in almost every data science role.

Search KDnuggets…

Latest Posts

- Forget Streamlit: Create an Interactive Data Science Dashboard in Excel in Minutes
- Go vs. Python for Modern Data Workflows: Need Help Deciding?
- 10 GitHub Repositories to Master Web Development in 2025
- Getting Started with Cassandra: Installation and Setup Guide
- The 7 Most Useful Jupyter Notebook Extensions for Data Scientists
- A Practical Guide to Multimodal Data Analytics

Top Posts

detailed and built on top of prerequisite knowledge I did not possess.

I spent time scouring the Internet for resources to better understand concepts like hypothesis testing and confidence intervals.

And after interviewing for multiple data science positions, I've found that most statistics interview questions followed a similar pattern.

In this article, I'm going to list 10 of the most popular statistics questions I've encountered in data science interviews, along with sample answers to these questions.

Question 1: What is a p-value?

Answer: Given that the null hypothesis is true, a p-value is the probability that you would see a result at least as extreme as the one observed.

P-values are typically calculated to determine whether the result of a statistical test is significant. In simple words, the p-value tells us whether there is enough evidence to reject the null hypothesis.

Question 2: Explain the concept of statistical power

Answer: If you were to run a statistical test to detect whether an effect is present, statistical power is the probability that the test will accurately detect the effect.

Here is a simple example to explain this:

Let's say we run an ad for a test group of 100 people and get 80 conversions.

The null hypothesis is that the ad had no effect on the number of conversions. In reality, however, the ad did have a significant impact on the amount of sales.

Statistical power is the probability that you would accurately reject the null hypothesis and actually detect the effect. A higher statistical power indicates that the test is better able to detect an effect if there is one.

Question 3: How would you describe confidence intervals to a non-technical stakeholder?

Let's use the same example as before, in which an ad is run for a sample size of 100 people and 80 conversions are obtained.

Instead of saying that the conversion rate is 80%, we would provide a range, since we don't know how the true population would behave. In other words, if we were to take an infinite number of samples, how many conversions would we see?

Machine Learning
MLOps
NLP
Programming
Python
Datasets
Events
Resources
Cheat Sheets
Recommendations
Tech Briefs
Advertise

rate will fall anywhere between 75% to 88%."

We use this range because we don't know how the total population will react, and can only generate an estimate based on our test group, which is just a sample.

Question 4: What is the difference between a parametric and non-parametric test?

A parametric test assumes that the dataset follows an underlying distribution. The most common assumption made when conducting a parametric test is that the data is normally distributed.

Examples of parametric tests include ANOVA, T-Test, F-Test and the Chi-squared test.

Non-parametric tests, however, don't make any assumptions about the dataset's distribution. If your dataset isn't normally distributed, or if it contains ranks or outliers, it is wise to choose a non-parametric test.

Question 5: What is the difference between covariance and correlation?

Covariance measures the direction of the linear relationship between variables. Correlation measures the strength and direction of this relationship.

While both correlation and covariance give you similar information about feature relationship, the main difference between them is scale.

Correlation ranges between -1 and +1. It is standardized, and easily allows you to understand whether there is a positive or negative relationship between features and how strong this effect is. On the other hand, covariance is displayed in the same units as the dependent and independent variables, which can make it slightly harder to interpret.

Question 6: How would you analyze and handle outliers in a dataset?

There are a few ways to detect outliers in the dataset.

- Visual methods: Outliers can be visually identified using charts like boxplots and scatterplots Points that are outside the whiskers of a boxplot are typically outliers. When using scatterplots, outliers can be detected as points that are far away from other data points in the visualization.
- Non-visual methods: One non-visual technique to detect outliers is the Z-Score. Z-Scores are computed by subtracting a value from the mean and dividing it by the standard

Blog
Top Posts
About
Topics
AI
Career Advice
Computer Vision
Data Engineering
Data Science
Language Models

outliers.

Question 7: Differentiate between a one-tailed and two-tailed test.

A one-tailed test checks whether there is a relationship or effect in a single direction. For example, after running an ad, you can use a one-tailed test to check for a positive impact, i.e. an increase in sales. This is a right-tailed test.

A two-tailed test examines the possibility of a relationship in both directions. For instance, if a new teaching style has been implemented in all public schools, a two-tailed test would assess whether there is a significant increase or decrease in scores.

Question 8: Given the following scenario, which statistical test would you choose to implement?

An online retailer want to evaluate the effectiveness of a new ad campaign. They collect daily sales data for 30 days before and after the ad was launched. The company wants to determine if the ad contributed to a significant difference in daily sales.

Options:

A) Chi-squared test

B) Paired t-test

C) One-way ANOVA

d) Independent samples t-test

Answer: To evaluate the effectiveness of a new ad campaign, we should use an paired t-test. A paired t-test is used to compare the means of two samples and check if a difference is statistically significant.
In this case, we are comparing sales before and after the ad was run, comparing a change in the same group of data, which is why we use a paired t-test instead of an independent samples t-test.

Question 9: What is a Chi-Square test of independence?

A Chi-Square test of independence is used to examine the relationship between observed and expected results. The null hypothesis (H0) of this test is that any observed difference between the features is purely due to chance.

In simple terms, this test can help us identify if the relationship between two categorical variables is due to chance, or whether there is a statistically significant association between them.

Blog
Top Posts
About
Topics
Career Advice
Computer Vision
Data Engineering
Data Science
Language Models

of independence.

Question 10: Explain the concept of regularization in regression models.

Regularization is a technique that is used to reduce overfitting by adding extra information to it, allowing models to adapt and generalize better to datasets that they haven't been trained on.

In regression, there are two commonly-used regularization techniques: ridge and lasso regression.

These are models that slightly change the error equation of the regression model by adding a penalty term to it.

In the case of ridge regression, a penalty term is multiplied by the sum of squared coefficients. This means that models with larger coefficients are penalized more. In lasso regression, a penalty term is multiplied by the sum of absolute coefficients.

While the primary objective of both methods is to shrink the size of coefficients while minimizing model error, ridge regression penalizes large coefficients more.

On the other hand, lasso regression applies a constant penalty to each coefficient, which means that coefficients can shrink to zero in some cases.

10 Statistics Questions to Ace Your Data Science Interview—Next Steps

If you’ve managed to follow along this far, congratulations!

You now have a strong grasp of the statistics questions asked in data science interviews.

As a next step, I recommend taking an online course to brush up on these concepts and put them into practice.

Here are some statistics learning resources I’ve found useful:

- StatQuest
- Krish Naik’s YouTube channel
- Statistical Learning on edX

The final course can be audited for free on edX, whereas the first two resources are YouTube channels that cover statistics and machine learning extensively.

Blog
Top Posts
About
Topics
AI
Career Advice
Computer Vision
Data Engineering
Data Science
Language Models

LinkedIn or check out her YouTube channel.

More On This Topic

- Top 7 Essential Cheat Sheets To Ace Your Data Science Interview
- 10 Cheat Sheets You Need To Ace Data Science Interview
- 7 Super Cheat Sheets You Need To Ace Machine Learning Interview
- 24 A/B Testing Interview Questions in Data Science Interviews and…
- Top 10 Advanced Data Science SQL Interview Questions You Must Know…
- 3 More SQL Aggregate Function Interview Questions for Data Science

<= Previous post Next post =>