Data Science Interview Questions and Answers

# Mathematics and Statistics

1. What is the Central Limit Theorem, and why is it important in statistics?

The Central Limit Theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size becomes large, regardless of the population's distribution. It is important because it allows for inference using normal distribution.

2. Explain the difference between population and sample.

A population includes all members of a defined group, while a sample is a subset of the population used to represent the group in statistical analysis.

3. What is probability and how is it calculated?

Probability is a measure of the likelihood of an event occurring, calculated as the number of favorable outcomes divided by the total number of possible outcomes.

4. What are the measures of central tendency, and when would you use each one?

The measures are mean, median, and mode. Mean is used for symmetric distributions, median for skewed distributions, and mode for categorical data.

5. Define variance and standard deviation.

Variance measures the average squared deviation from the mean, while standard deviation is the square root of variance, indicating data spread.

6. What is the significance of hypothesis testing in data science?

Hypothesis testing helps determine if there is enough evidence to support a specific claim about a dataset.

7. Explain the p-value and its significance in hypothesis testing.

The p-value indicates the probability of observing the data assuming the null hypothesis is true. A low p-value suggests rejecting the null hypothesis.

8. What is a normal distribution, and why is it important in statistics?

A normal distribution is a symmetric, bell-shaped distribution. It is important because many statistical tests assume normality.

9. Describe the differences between a Z-score and a T-score.

Z-scores are used when population variance is known and sample size is large; T-scores are used when population variance is unknown and sample size is small.

10. What is correlation, and how is it measured?

Correlation measures the strength and direction of a linear relationship between two variables, typically using Pearson’s correlation coefficient.

11. What is the difference between covariance and correlation?

Covariance indicates the direction of a linear relationship, while correlation standardizes this measure to a range between -1 and 1.

12. What is the law of large numbers?

It states that as the sample size increases, the sample mean will converge to the population mean.