Statistics Interview Questions

(Practice Project)

PW SKILLS

Easy

1. What are descriptive statistics?

Ans: Descriptive statistics summarizes and organizes the characteristics of a dataset. It includes measures of central tendency, measures of variability (dispersion), and graphical representations.

2. Explain the difference between the mean and median.

Ans: The mean is the average of all data points, while the median is the middle value when the data points are sorted. The median is less affected by outliers and skewed data than the mean.

3. How do you calculate the standard deviation?

Ans: Standard deviation is calculated by taking the square root of the average of the squared deviations from the mean.

4. What is skewness and how is it calculated?

Ans: Skewness measures the asymmetry of the probability distribution of a real-valued random variable. Positive skewness indicates a distribution with a long tail on the right side, while negative skewness indicates a long tail on the left side. It can be calculated using: from scipy.stats import
skew

data = [1, 2, 3, 4, 5, 6, 7, 8, 100]
skewness = skew(data)

5. What are the three main measures of central tendency?

Ans: The three main measures of central tendency are the mean, median, and mode.

Medium

6. Explain the concept of kurtosis.

Ans: Kurtosis measures the "tailedness" of the probability distribution. High kurtosis indicates heavy tails and sharp peaks, while low kurtosis indicates light tails and flatter peaks.

7. What are the advantages and disadvantages of using the mean as a measure of central tendency?

Ans: Advantages: The mean takes all values into account, providing a comprehensive measure. It is useful for further statistical analysis.Disadvantages: The mean is sensitive to outliers and skewed data, which can distort the measure.

8. How do you decide which measure of central tendency to use?

Ans: The choice depends on the nature of the data and the specific context:

- Use the mean for symmetrical distributions without outliers.
- Use the median for skewed distributions or when outliers are present.
- Use the mode for categorical data or to find the most common value in a dataset

9. How can central tendency measures be misleading in a skewed distribution?

Ans: In a skewed distribution, the mean can be pulled toward the tail, giving a misleading representation of the center. The median provides a better central value in such cases. For example, in a right-skewed distribution of incomes, the mean may be higher than the median, suggesting a higher average income than most people actually earn.

10. What are the measures of dispersion?

Ans: Measures of dispersion describe the spread or variability within a set of data. Common measures include range, variance, standard deviation, and interquartile range (IQR).

11. Explain the concept of skewness and how it affects measures of dispersion.

Ans: Skewness describes the asymmetry of the data distribution. In a skewed distribution, the mean and standard deviation may be misleading. For right-skewed data, the mean is greater than the median, and for left-skewed data, the mean is less than the median. Dispersion measures like IQR are less affected by skewness.

Hard

12. Explain the concept of a probability distribution.

Ans: A probability distribution describes how the probabilities are distributed over the values of the random variable. For discrete random variables, it is defined by the probability mass function (PMF), and for continuous random variables, it is defined by the probability density function (PDF).

13. What is the normal distribution, and why is it important?

Ans: The normal distribution is a continuous probability distribution characterized by a symmetric, bell-shaped curve. It is important because it describes many natural phenomena, and the Central Limit Theorem states that the means of large samples of any distribution will be normally distributed.

14. Describe a situation where the binomial distribution would be more appropriate than the normal distribution.

Ans: The binomial distribution is more appropriate in scenarios with a fixed number of independent trials and two possible outcomes, such as testing the success rate of a new drug in a fixed number of patients or the pass rate of students in an exam.

15. What is inferential statistics?

Ans: Inferential statistics involves making predictions or inferences about a population based on a sample of data drawn from that population. It includes hypothesis testing, confidence intervals, and regression analysis.

16. What is correlation?
Ans: Correlation is a statistical measure that describes the extent to which two variables are linearly related. It is represented by the correlation coefficient, which ranges from -1 to 1.

17. How is covariance different from correlation?

Ans: Covariance measures the direction of the linear relationship between two variables but not the strength or scale. Correlation, on the other hand, standardizes the covariance by the product of the standard deviations of the variables, providing a dimensionless measure that ranges from -1 to 1, indicating both the strength and direction of the relationship.

18. How does scaling of data affect covariance?

Ans: Covariance is sensitive to the scale of the data. If the units of the variables are changed, the covariance will also change. For example, if one variable is measured in dollars and another in cents, the covariance will be affected by the unit conversion. This is why correlation, which is a scaled version of covariance, is often used for comparison purposes.