Top 20 Interview Questions in Descriptive Statistics for Data Science

Data Science Student

September 19, 2024

1 Questions and Answers

1. What is descriptive statistics, and how is it different from inferential statistics?

Answer: Descriptive statistics involves summarizing and describing the features of a dataset, such as through measures of central tendency (mean, median, mode) and dispersion (range, variance, standard deviation). It does not make predictions beyond the data at hand. Inferential statistics, on the other hand, uses samples to infer or predict characteristics of the larger population, employing methods such as hypothesis testing and confidence intervals.

2. What are the key measures of central tendency in descriptive statistics?

Answer: The key measures of central tendency are:

- **Mean:** The average of all data points.
- **Median:** The middle value in an ordered dataset.
- **Mode:** The value that appears most frequently in the dataset.

3. When would you use the median over the mean?

Answer: The median is preferred over the mean when the dataset contains **outliers** or is **skewed**. The median is not affected by extreme values, whereas the mean is sensitive to outliers, which can distort the measure of central tendency.

4. What is the mode, and in what scenarios is it most useful?

Answer: The **mode** is the most frequently occurring value in a dataset. It is especially useful when analyzing **categorical data**, where you want to identify the most common category or value.

5. How do you calculate the range, and what does it signify?

Answer: The range is calculated as the difference between the maximum and minimum values in a dataset:

$$\text{Range} = \text{Max} - \text{Min}$$

It signifies the **spread** of the data, providing a quick measure of variability.

6. What is the interquartile range (IQR), and why is it important?

Answer: The **interquartile range (IQR)** is the difference between the 75th percentile (Q3) and the 25th percentile (Q1):

$$\text{IQR} = Q_3 - Q_1$$

It is important because it provides a robust measure of variability, less influenced by outliers than the range.

7. What is variance, and how is it calculated?

Answer: Variance measures the overall spread of data points from the mean. For a population, it is calculated as:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

For a sample, the formula uses $n-1$ in the denominator to provide an unbiased estimate:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

8. What is standard deviation, and how is it related to variance?

Answer: The **standard deviation** is the square root of the variance:

$$\sigma = \sqrt{\sigma^2}$$

It provides a measure of how much the data deviates from the mean, in the same units as the data.

9. Why is the sample variance divided by $n-1$ instead of n?

Answer: Dividing by $n-1$ corrects for the bias in the estimation of the population variance from a sample. This adjustment, known as **Bessel's correction**, gives an unbiased estimate of the population variance.

10. What is the empirical rule (68-95-99.7 rule) in statistics?

Answer: The **empirical rule** applies to normally distributed data. It states that:

- 68% of the data falls within one standard deviation of the mean.
- 95% falls within two standard deviations.
- 99.7% falls within three standard deviations.

11. What is a z-score, and how is it calculated?

Answer: A **z-score** represents how many standard deviations a data point is from the mean:

$$z = \frac{x - \mu}{\sigma}$$

A z-score of 0 indicates that the data point is exactly at the mean.

12. What is skewness in a dataset?

Answer: Skewness measures the asymmetry of a dataset's distribution. A dataset is **positively skewed** if its tail is longer on the right, and **negatively skewed** if the tail is longer on the left.

13. What is kurtosis, and what does it tell you about a dataset?

Answer: Kurtosis describes the "tailedness" of a distribution:

- **High kurtosis** indicates more data in the tails, meaning more extreme values.
- **Low kurtosis** indicates less data in the tails, meaning fewer outliers.

14. What is the difference between population and sample statistics?

Answer: - **Population statistics** describe the entire dataset and use population parameters like μ (population mean) and σ (population standard deviation). - **Sample statistics** describe a subset of the population and use sample estimates like $\bar{x}$ (sample mean) and s (sample standard deviation).

15. Why is the median a robust measure in the presence of outliers?

Answer: The median is robust because it is the middle value of the dataset and does not depend on extreme values (outliers). It is a better measure of central tendency when the data is skewed or contains outliers.

16. What is the difference between univariate and bivariate analysis?

Answer: - **Univariate analysis** focuses on analyzing a single variable (e.g., using mean, median, mode). - **Bivariate analysis** examines the relationship between two variables (e.g., using correlation or regression analysis).

17. What is the purpose of a box plot, and what does it show?

Answer: A **box plot** displays the distribution of a dataset through five summary statistics: minimum, Q1, median, Q3, and maximum. It also shows outliers and the spread of the middle 50% of the data (the interquartile range, IQR).

18. What is a percentile, and how does it differ from a quartile?

Answer: A **percentile** indicates the value below which a given percentage of the data falls. A **quartile** divides the data into four equal parts:

- **Q1 (25th percentile):** 25% of the data is below this value.
- **Q2 (50th percentile):** The median.
- **Q3 (75th percentile):** 75% of the data is below this value.

19. What is the purpose of data normalization?

Answer: Data normalization scales the values in a dataset to a standard range, usually between 0 and 1. It ensures that no feature dominates the others in algorithms like k-means clustering or neural networks.

20. What is an outlier, and how can you detect it?

Answer: An **outlier** is a data point significantly different from the other points in the dataset. Outliers can be detected using:

- The **IQR method:** Data points that are more than 1.5 times the IQR from Q1 or Q3.
- The **z-score method:** Data points with z-scores greater than 3 or less than -3.