<https://chatgpt.com/share/b58fb359-9b15-44f5-8d31-8c6f888b6ecd>

 50 questions covering various topics in descriptive statistics and probability distributions, which are fundamental to data science:

Mean, Median, Mode

1. What is the mean of the dataset: [10, 15, 12, 18, 20]?
2. Calculate the median of the dataset: [5, 7, 12, 13, 15, 16, 20].
3. Determine the mode of the dataset: [6, 8, 8, 10, 10, 10, 12].
4. How do you interpret the mean of a dataset in terms of central tendency?
5. If a dataset is [4, 9, 11, 15, 17, 20], what is the median if an additional value 10 is added to the dataset?

Variance and Standard Deviation

1. Calculate the variance of the dataset: [2, 4, 4, 4, 5, 5, 7, 9].
2. What is the standard deviation of the dataset: [1, 3, 5, 7, 9]?
3. Explain the relationship between variance and standard deviation.
4. Why is the standard deviation preferred over variance when comparing variability?
5. Compute the variance and standard deviation of the dataset: [12, 15, 14, 16, 18].

Correlation and Covariance

1. Define the correlation coefficient and its range.
2. Calculate the correlation coefficient between [X: 1, 2, 3, 4, 5] and [Y: 2, 4, 5, 4, 5].
3. What is covariance and how is it different from correlation?
4. Compute the covariance of the datasets [X: 1, 2, 3, 4, 5] and [Y: 3, 6, 7, 8, 9].
5. Why is correlation a better measure of the relationship between two variables compared to covariance?

Skewness and Kurtosis

1. Define skewness and describe the difference between positive and negative skewness.
2. Given the dataset [3, 5, 6, 8, 9, 15, 21], determine if it is positively or negatively skewed.
3. What is kurtosis and skewness what does it tell us about the distribution of data

Probability Distributions

1. Define a probability distribution and its properties.
2. What is the difference between a discrete and a continuous probability distribution?
3. Provide an example of a real-world scenario that can be modeled using a normal distribution.
4. Explain the binomial distribution and give an example of its application.
5. Describe the Poisson distribution and its use cases.

Probability Density Function (PDF) and Cumulative Distribution Function (CDF)

1. What is a probability density function (PDF) and how is it used?
2. Explain the cumulative distribution function (CDF) and its significance.
3. How is the CDF related to the PDF?
4. For a continuous random variable, describe how you would use the PDF to find probabilities.
5. Given a normal distribution with mean 0 and standard deviation 1, calculate the CDF at (z = 1.96).

Probability Mass Function (PMF)

1. Define the probability mass function (PMF) and its properties.
2. How is the PMF used in the context of discrete random variables?
3. Provide an example of a PMF for a fair six-sided die.
4. Explain the relationship between the PMF and the cumulative distribution function (CDF) for discrete random variables.
5. Calculate the PMF for the number of heads in two coin tosses.

Practical Applications

1. Describe a scenario in data science where calculating the mean would be crucial.
2. Explain why understanding the variance is important when analyzing a dataset.
3. How can correlation analysis be used in predicting sales based on advertising spend?
4. Describe how skewness can affect the choice of statistical methods in data analysis.
5. Why is it important to consider kurtosis when analyzing financial returns?

Advanced Concepts

1. What is the significance of the central limit theorem in probability distributions?
2. How does the law of large numbers apply to empirical data analysis?
3. Explain the difference between a population and a sample in statistics.
4. What is the role of standard error in inferential statistics?
5. How would you interpret a z-score in the context of a normal distribution?

Visualization

1. Create a histogram for the dataset: [10, 15, 15, 20, 25, 30, 35, 40].
2. Draw a box plot for the dataset: [5, 6, 7, 8, 9, 10, 15, 20] and identify any outliers.
3. Explain how a scatter plot can be used to identify correlation between two variables.
4. What information can be obtained from a QQ plot?
5. Describe the use of a heatmap in visualizing correlation matrices.

#### **Section 1: Understanding Data and Visualization**

**1.1. Definitions and Concepts**

1. **Question:** Define data. How is data different from information?

**Expected Answer:** Data refers to raw facts and figures that are unprocessed, while information is data that has been processed or interpreted to have meaning.

1. **Question:** What is a line chart, and when would you use it?

**Expected Answer:** A line chart displays data points connected by straight lines. It is used to show trends over time or continuous data.

1. **Question:** Explain the difference between a scatter chart and a line chart.

**Expected Answer:** A scatter chart displays individual data points and is used to observe relationships or correlations between two variables, while a line chart connects data points with lines to show trends.

1. **Question:** What is the purpose of a bar graph? How does it differ from a histogram?

**Expected Answer:** A bar graph is used to compare quantities across different categories, while a histogram shows the distribution of a continuous variable by grouping data into bins.

1. **Question:** Describe when you would use a pie chart and its potential limitations.

**Expected Answer:** A pie chart shows proportions or percentages of a whole. It is best used for categorical data with limited categories, but can be less effective with many categories or small differences.

**1.2. Practical Assignment**

* **Assignment:** Given a dataset containing monthly sales figures for different product categories, create a line chart to visualize sales trends over the year. Additionally, create a bar graph comparing sales across product categories.

#### **Section 2: Characterizing Data with Descriptive Statistics**

**2.1. Central Tendency and Dispersion**

1. **Question:** Define and differentiate between the mean, median, and mode.

**Expected Answer:** Mean is the average value, median is the middle value when data is sorted, and mode is the most frequent value.

1. **Question:** What are measures of dispersion, and why are they important?

**Expected Answer:** Measures of dispersion include range, variance, and standard deviation. They describe the spread or variability of the data, which helps understand the consistency of the data.

**2.2. Probability Basics and Distributions**

1. **Question:** Explain the concept of skewness in a dataset.

**Expected Answer:** Skewness measures the asymmetry of a data distribution. Positive skew indicates a longer tail on the right, while negative skew indicates a longer tail on the left.

1. **Question:** What is a uniform distribution?

**Expected Answer:** A uniform distribution is one where all outcomes are equally likely, and each value within a range has the same probability of occurring.

1. **Question:** Describe the binomial distribution and provide an example of its application.

**Expected Answer:** The binomial distribution describes the number of successes in a fixed number of independent Bernoulli trials. An example is flipping a coin multiple times and counting the number of heads.

1. **Question:** What is a standard normal distribution?

**Expected Answer:** A standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. It is used as a reference for comparing different data distributions.

**2.3. Practical Assignment**

* **Assignment:** Given a dataset with exam scores, calculate the mean, median, mode, variance, and standard deviation. Create a histogram to visualize the distribution and determine the skewness of the data.

#### **Section 3: Predicting from Data with Inferential Statistics**

**3.1. Sampling and Hypothesis Testing**

1. **Question:** What is random sampling and how does it differ from stratified sampling?

**Expected Answer:** Random sampling involves selecting a sample where each member of the population has an equal chance of being chosen. Stratified sampling divides the population into subgroups and samples from each subgroup to ensure representation.

1. **Question:** Explain the Central Limit Theorem and its significance.

**Expected Answer:** The Central Limit Theorem states that the distribution of the sample mean will be approximately normal if the sample size is large enough, regardless of the population distribution. It is significant for making inferences about population parameters.

1. **Question:** What is hypothesis testing, and how do you interpret p-values?

**Expected Answer:** Hypothesis testing is a method to determine if there is enough evidence to reject a null hypothesis. A p-value indicates the probability of obtaining the observed results if the null hypothesis is true; a low p-value suggests strong evidence against the null hypothesis.

1. **Question:** Define Type 1 and Type 2 errors.

**Expected Answer:** Type 1 error occurs when a true null hypothesis is incorrectly rejected (false positive), while Type 2 error occurs when a false null hypothesis is not rejected (false negative).

**3.2. Practical Assignment**

* **Assignment:** Given a dataset, perform a hypothesis test to determine if there is a significant difference between the means of two groups. Report the p-value and discuss the potential Type 1 and Type 2 errors.

**Questions Set Statistics**

1. Define data. How is data different from information?
2. What is a line chart, and when would you use it?
3. Explain the difference between a scatter chart and a line chart.
4. What is the purpose of a bar graph? How does it differ from a histogram?
5. Describe when you would use a pie chart and its potential limitations.
6. Define and differentiate between the mean, median, and mode.
7. What are measures of dispersion, and why are they important?
8. Explain the concept of skewness in a dataset.
9. What is a uniform distribution?
10. Describe the binomial distribution and provide an example of its application.
11. What is a standard normal distribution?
12. What is random sampling and how does it differ from stratified sampling?
13. Explain the Central Limit Theorem and its significance.
14. What is hypothesis testing, and how do you interpret p-values?
15. Define Type 1 and Type 2 errors.
16. Describe different data collection techniques and their pros and cons.
17. Explain the process of data cleaning and why it is crucial.
18. Explain the central limit theorem
19. How would you describe a 'p-value
20. What is sampling?
21. What is linear regression?
22. What is the normal distribution?
23. What is bias in data science?
24. Where is inferential statistics used?
25. Difference between population and sample
26. What does standard deviation mean?
27. What is a statistical interaction?
28. What is an outlier?
29. What is selection bias?
30. What is the confidence interval?
31. What is the Pareto Principle?
32. Describe hypothesis testing
33. What are descriptive statistics?
34. What are left-skewed and right-skewed distributions
35. What is Bessel's correction?
36. Exploratory data analysis
37. What is root cause analysis?
38. What is the assumption of normality?
39. What is the binomial distribution formula?
40. How do data scientists use statistics
41. Explain the central limit theorem and give examples of when you can use it in a real-world problem?
42. Briefly explain the A/B testing and its application? What are some common pitfalls encountered in A/B testing?
43. Describe briefly the hypothesis testing and p-value in layman’s term? And give a practical application for them?
44. Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?
45. What is the meaning of selection bias and how to avoid it?
46. Explain the long-tailed distribution and provide three examples of relevant phenomena that have long tails. Why are they important in classification and regression problems?
47. What is the meaning of KPI in statistics
48. Say you flip a coin 10 times and observe only one head. What would be the null hypothesis and p-value for testing whether the coin is fair or not?
49. You are testing hundreds of hypotheses, each with a t-test. What considerations would you take into account when doing this?
50. What general conditions must be satisfied for the central limit theorem to hold?
51. What is skewness discuss two methods to measure it?
52. You sample from a uniform distribution [0, d] n times. What is your best estimate of d?
53. Discuss the Chi-square, ANOVA, and t-test
54. Say you have two subsets of a dataset for which you know their means and standard deviations. How do you calculate the blended mean and standard deviation of the total dataset? Can you extend it to K subsets?
55. What is the relationship between the significance level and the confidence level in Statistics?
56. What is the Law of Large Numbers in statistics and how it can be used in data science ?
57. What is the difference between a confidence interval and a prediction interval, and how do you calculate them?
58. What are the differences between the z-test and t-test?
59. When to use a z-test Vs a t-test?
60. Given a specific dataset, how do you calculate t-statistic or z-statistics?

<https://intellipaat.com/blog/interview-question/statistics-interview-questions/>

<https://365datascience.com/career-advice/job-interview-tips/probability-and-statistics-interview-questions-for-data-scientists/>

<https://github.com/youssefHosni/Data-Science-Interview-Questions-Answers/blob/main/Statistics%20Interview%20Questions%20%26%20Answers%20for%20Data%20Scientists.md>