

STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log-normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

6. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

8. Normalized data are centered at _____ and have units equal to standard deviations of the original data.

- a) 0
- b) 5
- c) 1
- d) 10

9. Which of the following statement is incorrect with respect to outliers?

- a) Outliers can have varying degrees of influence
- b) Outliers can be the result of spurious or real processes
- c) Outliers cannot conform to the regression relationship
- d) None of the mentioned

10. What do you understand by the term Normal Distribution?

Ans:- Normal distribution, also known as the Gaussian distribution, is a probability distribution that forms a symmetric, bell-shaped curve when graphed. It's characterized by two parameters: the mean (average) and the standard deviation. In a normal distribution, most data points cluster around the mean, with fewer points further away towards the tails of the curve. This distribution is widely used in statistics and probability theory to model various phenomena in fields such as natural sciences, social sciences, finance, and engineering.

11. How do you handle missing data? What imputation techniques do you recommend?

Ans: - Handling missing data involves deciding what to do with the gaps in your dataset. Here's a simple approach:

1. Find Missing Data: Identify where data is missing.
2. Decide on Imputation: Choose a method to fill in the missing values.
3. Common Methods:
 - Mean or Median: If we're missing numbers, we can calculate the average (mean) or middle value (median) of the numbers we do have, and use that for the missing ones.

- **Most Common:** If we're missing categories (like types of cars), we can look at the ones we already know about and use the most common one for the missing data.
- **Last Observation:** If our data follows a pattern over time, like temperatures or stock prices, we can just copy the last known value for anything we're missing.

4. **Evaluate:** Check if imputation affects your analysis.

5. **Document:** Keep track of what you did and why you chose that method.

12. What is A/B testing?

Ans: - A/B testing, also known as split testing, is a method used to compare two versions of something to see which one performs better. It's commonly used in marketing, web design, and product development to make data-driven decisions.

13. Is mean imputation of missing data acceptable practice?

Ans: - Mean imputation is a straightforward method for handling missing data where the missing values are replaced with the mean of the observed data for that variable. While it's simple and easy to implement, it may not always be the best approach as it can introduce bias and underestimate variability, especially if missingness is not random or if the data is heavily skewed. Therefore, while mean imputation can be acceptable in certain situations, it's important to consider its limitations and potential impact on the analysis.

14. What is linear regression in statistics?

Ans: - linear regression is a way to understand and predict how one variable changes as another variable changes. It's like drawing a straight line through a bunch of points on a graph to show the overall trend. This helps us see how two things are related and make predictions based on that relationship.

15. What are the various branches of statistics?

Ans: - Statistics is a broad field with various branches that focus on different aspects of data analysis and interpretation. Some of the main branches of statistics include:

- **Descriptive Statistics:** Summarizes data using measures like mean and standard deviation.
- **Inferential Statistics:** Draws conclusions about populations based on samples.
- **Probability Theory:** Studies random events and uncertainty.
- **Biostatistics:** Analyzes data in biology and health sciences.
- **Econometrics:** Applies statistics to economic data for analysis and forecasting.
- **Spatial Statistics:** Analyzes data with spatial components, like maps.
- **Time Series Analysis:** Studies data collected over time for trends and patterns.

- **Multivariate Statistics:** Analyzes datasets with multiple variables simultaneously.