STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

a) True

b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

8. 4. 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

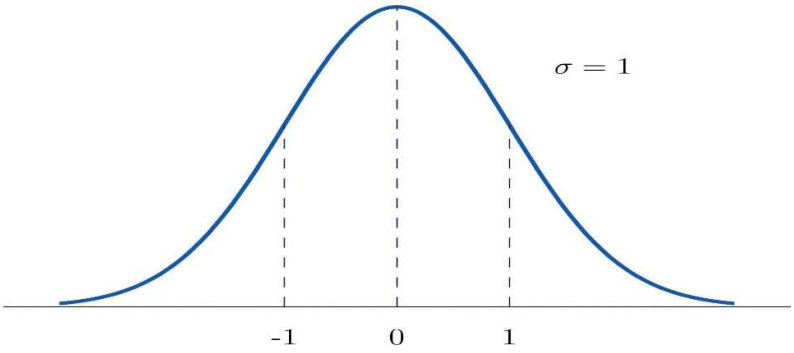
d) None of the mentioned

WORKSHEET

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

A Normal Distribution is a bell-shaped probability distribution that is symmetric about the mean. Most of the data points lie close to the mean, with fewer and fewer points further away from the mean, creating the characteristic bell shape. The normal distribution is characterized by two parameters: the mean (which determines the center of the distribution) and the standard deviation (which measures the spread or width of the distribution).



Normal Distribution Curve

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

Handling missing data depends on the nature of the data and the extent of the missing data.

To handle missing data, you can;

a. Delete which is to remove rows or columns with missing data,
b. Replace with either the mean or median values of the respective feature,
c. Imputation of the mode to replace categorical values.
d. Use K-nearest Neighbours (KNN) to predict missing values based on nearest neighbour in the data set.

12. What is A/B testing?

A/B testing is a statistical method used to compare two versions (A and B) of a variable (such as a webpage, email, or product feature) to determine which one performs better. It involves randomly splitting the sample into two groups: one group sees version A, and the other sees version B. The performance of both versions is then compared using a statistical test to determine if the difference is significant or not significant.

13. Is mean imputation of missing data acceptable practice?

Not really because it reduces the variance in the dataset, potentially underestimating the variability of the data although it s simple and quick. it is generally considered a **basic** and **suboptimal** imputation technique, especially for datasets with significant amounts of missing data.

14. What is linear regression in statistics?

is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to fit a linear equation to the data such that the sum of squared differences between the observed values and the values predicted by the linear equation is minimized.

The basic formular of a linear regression model is:

Y= a+b*x+e

Where:

a=Intercept

b=slope

c=error

15. What are the various branches of statistics?

a. **Descriptive Statistics**: Involves summarizing and describing the features of a dataset through measures like mean, median, mode, variance, and graphs (e.g., histograms, pie charts).
b. **Inferential Statistics**: Focuses on making predictions or inferences about a population based on a sample of data. It includes hypothesis testing, confidence intervals, and regression analysis.
c. **Probability Theory**: The mathematical foundation of statistics, dealing with the likelihood of events occurring.
d. **Predictive Analytics**: Uses statistical models and machine learning techniques to predict future outcomes based on current and historical data.
e. **Exploratory Data Analysis (EDA)**: Involves analysing data sets to summarize their main characteristics, often with visual methods.
f. **Bayesian Statistics**: A subfield of statistics in which probability expresses a degree of belief in an event, based on prior knowledge updated by new evidence.