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

d) None of the mentioned

The correct answer is:
a) True
A **Bernoulli random variable** is a discrete random variable that takes only two possible values, typically **1** (success) and **0** (failure), with a certain probability. The probability of success is denoted by **p**, and the probability of failure is **1 - p**.

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

a) Central Limit Theorem
The **Central Limit Theorem (CLT)** states that the distribution of the sample means (averages) of independent and identically distributed (iid) random variables, when properly normalized, tends to become a standard normal distribution (i.e., a normal distribution with mean 0 and variance 1) as the sample size increases, regardless of the original distribution of the data.

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

Answer: b) Modeling bounded count data
The Poisson distribution is typically used to model counts of events occurring in a fixed interval of time or space, where the counts are not bounded. However, it is not well-suited for **bounded count data** (data with a fixed upper limit).

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

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

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

c) Poisson
Poisson random variables are commonly used to model rates, specifically the number of events occurring within a fixed period of time or space, such as the number of calls received at a call center per hour.

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

Answer: b) False
Replacing the standard error with an estimated value generally does not change the application of the **Central Limit Theorem** (CLT), as the CLT still applies even when parameters are estimated from the data.

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

Answer: b) Hypothesis
Hypothesis testing is a statistical method concerned with making decisions based on data, specifically to test assumptions or claims about population parameters.

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

Answer: a) 0
Normalized (or standardized) data have a mean of 0 and a standard deviation of 1.

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

Answer: d) None of the mentioned
All of the other statements are correct: outliers can have varying influence, can result from spurious or real processes, and they can deviate from the regression relationship.

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 continuous probability distribution characterized by its bell-shaped curve, symmetric about the mean. The mean, median, and mode of a normal distribution are equal. It is fully described by two parameters: the mean (which determines the center) and the standard deviation (which determines the spread or width of the curve). The total area under the curve is 1, and it represents the probability distribution of a wide range of phenomena in nature, economics, and other fields.

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 missingness.
Some common techniques include:

Removing missing data: If the percentage of missing values is small, deleting those rows or columns can be effective.

Mean/Median/Mode imputation: Replacing missing values with the mean, median, or mode of the data.

Forward/Backward fill: In time series data, missing values can be filled using nearby data points.

Predictive modeling: Using algorithms like k-nearest neighbors (KNN) or regression models to predict missing values.

Multiple imputation: Replacing missing values multiple times to account for uncertainty in the imputed values.

The best technique depends on the dataset and should consider the potential bias introduced by the imputation.

11. What is A/B testing?

A/B testing is a statistical method used to compare two versions (A and B) of a variable to determine which one performs better. It is commonly used in fields like marketing and product development to assess changes such as website designs, product features, or ad campaigns. The two groups (A and B) are exposed to different treatments, and the results (usually in terms of user

engagement or performance) are statistically analyzed to see which variant yields better results.

12. Is mean imputation of missing data acceptable practice?

Mean imputation is a simple and commonly used technique to handle missing data by replacing missing values with the mean of the available data. However, it is not always the best practice because it can distort the variance of the data and reduce statistical power. It is acceptable in certain scenarios where the percentage of missing data is small and the data distribution is relatively normal. More advanced methods like multiple imputation or predictive modeling are often preferred.

13. What is linear regression in statistics?

Linear regression is a statistical method used to model the relationship between a dependent variable (often called the outcome or response) and one or more independent variables (predictors). The relationship is assumed to be linear, and the model seeks to find the line (in simple linear regression) or plane (in multiple regression) that best fits the data by minimizing the differences between the observed and predicted values (usually through least squares estimation).

14. What are the various branches of statistics?

The various branches of statistics include:

Descriptive statistics: Summarizing and describing the features of a dataset using measures such as mean, median, mode, and standard deviation.

Inferential statistics: Making predictions or inferences about a population based on a sample, using techniques like hypothesis testing, confidence intervals, and regression analysis.

Probability theory: Studying randomness and uncertainty, often forming the theoretical basis for inferential statistics.

Exploratory Data Analysis (EDA): Using statistical tools to explore datasets and uncover patterns, anomalies, and relationships.

Predictive analytics: Using statistical models and algorithms to forecast future trends based on historical data.

Bayesian statistics: A subfield of statistics that involves updating the probability of a hypothesis as more evidence or information becomes available.

Non-parametric statistics: Methods that do not assume a specific probability distribution in the data, useful when the assumptions of parametric tests are violated.