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 Answer
   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 Answer
   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 Answer
   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 Answer
   d) All of the mentioned
5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson Answer
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False Answer
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis Answer
   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 Answer
   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 Answer
   d) None of the mentioned

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

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

Normal distribution is a statistical distribution that is symmetric and bell-shaped, with the highest frequency of observations at the mean value and progressively fewer observations farther away from the mean in either direction. Its also known as Gaussian distribution or bell curve,

Its characterized by two parameters: the mean (μ) and the standard deviation (σ). The mean represents the center or average of the distribution, while the standard deviation represents the spread or variability of the distribution. The probability density function (PDF) of the normal distribution is given by the formula:

$f(x) = (1/\sigma\sqrt{2\pi}) * e^{(-(x-\mu)^2/2\sigma^2)}$

where x is the random variable, σ is the standard deviation, μ is the mean, e is the base of the natural logarithm, and π is the mathematical constant pi.

11. How do you handle missing data? What imputation techniques do you recommend?
Various techniques that can be used to handle missing data, are:

Complete case analysis: Involves excluding observations with missing values, which can lead to loss of information and reduced sample size.

Mean/median imputation: This is replacing missing values with the mean or median of the available values for the variable. While this method is simple to implement, it can lead to biased estimates and distorted variance estimates.

Multiple imputation: Its creating multiple plausible imputations for the missing values based on the observed data and modeling the relationships between the variables. This method is more robust than mean/median imputation and can provide unbiased estimates and standard errors.

Maximum likelihood estimation: Using a statistical model to estimate the missing values based on the available data. This method can provide unbiased estimates and can be used for both continuous and categorical variables.

Regression imputation: This involves using regression models to estimate the missing values based on the relationships between the variables. This method can provide accurate estimates if the relationships between the variables are well-understood.

I will recommend multiple imputation, as it is a flexible and robust method that can handle different types of missing data and can provide accurate estimates and standard errors. However, it is important to carefully evaluate the assumptions and limitations of any imputation method used and to report the results appropriately.

12. What is A/B testing?
Also known as split testing, is a statistical technique used to compare two versions of a product, service, or marketing campaign to determine which one performs better. It involves randomly assigning participants to one of two groups: group A, which receives the control or original version, and group B, which receives the experimental or modified version. The performance of the two versions is then compared based on a predefined metric, such as click-through rate, conversion rate, or revenue.

It is usualy used in website design, email marketing, and product development to improve user experience and increase conversion rates. It allows businesses to test different design elements, messaging, pricing, and other variables to determine which ones are most effective in achieving their goals.

13. Is mean imputation of missing data acceptable practice?

    It may be acceptable in certain circumstances, such as when the proportion of missing values is small and the data is missing completely at random (MCAR) or MAR, it's generally not recommended as a standalone method for handling missing data. Instead, multiple imputation or other more sophisticated methods should be used, as they can provide more accurate and reliable estimates and account for the uncertainty associated with missing data.

14. What is linear regression in statistics?

    Linear regression is a statistical technique used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The goal of linear regression is to find the best linear equation that describes the relationship between the variables

15. What are the various branches of statistics?
    Various branches are areas which knowledge of statistics are applied and include:

    Descriptive statistics: It deals with the organization, summarization, and description of data using measures such as mean, median, mode, range, variance, and standard deviation.

    Inferential statistics: It makes inferences and predictions about a population based on a sample of data. It includes hypothesis testing, confidence intervals, and regression analysis.

    Biostatistics: Used for statistical methods in the field of medicine, public health, and biology. It includes clinical trials, epidemiology, and survival analysis.

    Econometrics: Its a branch of statistics that deals with the application of statistical methods in economics and finance. It includes time series analysis, forecasting, and regression analysis.

    Psychometrics: It involves the development and validation of psychological tests and measures. It includes item analysis, factor analysis, and reliability analysis.

    Data science: The use of statistical and computational methods to extract insights and knowledge from data. It includes data mining, machine learning, and big data analytics.

    Statistical genetics: For the analysis of genetic data using statistical methods. It includes genome-wide association studies, linkage analysis, and haplotype analysis.

    Quality control: This branch of statistics deals with the use of statistical methods to ensure the quality and consistency of products and services. It includes process control, statistical process control, and acceptance sampling.

    These are just some of the major branches of statistics, and there are many other subfields and specialized areas within the discipline.