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

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

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

Answer: The normal distribution, also known as the Gaussian distribution or bell curve, is a probability distribution that describes the distribution of data that is symmetric around the mean. It is defined by its mean (average) and standard deviation (spread) and is often used to model real-world data. It is a continuous probability distribution that is defined by a probability density function, which is characterized by a bell-shaped curve. The normal distribution is important in statistics and is used in many areas of science and engineering, including finance, economics, and the natural sciences.

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

Answer: Handling missing data is an important step in data preprocessing, and it can have a significant impact on the results of an analysis. There are several techniques for handling missing data, including:
1. Deletion: This technique involves removing observations with missing data from the dataset. This is simple and easy to implement, but it can lead to a loss of information and can be problematic if the missing data is not missing completely at random (MCAR).
2. Mean/mode imputation: This technique involves replacing missing values with the mean or mode of the non-missing values for that variable. This is also simple and easy to implement, but it can lead to biased estimates if the missing data is not MCAR.
3. Predictive imputation: This technique involves using a predictive model to estimate missing values based on the other variables in the dataset. This can be more accurate than mean/mode imputation, but it can also lead to biased estimates if the missing data is not MCAR.
4. Multiple imputation: This technique creates multiple imputed datasets and combines them using statistical methods to account for the uncertainty in the imputed values. It is more robust to missing data, but it can be more complex and computationally intensive.
5. Hot-Deck Imputation: This method imputes the missing value with the value from a similar case in the dataset.
6. Cold-Deck Imputation: This method creates a separate dataset that is used to impute missing values.
7. Expectation-Maximization (EM) imputation: This method estimates missing values by using the Expectation-Maximization algorithm.
8. Regression imputation: This method uses a regression model to predict missing values based on other variables in the dataset.
9. Interpolation: This method uses the values of the neighboring observations to estimate missing values.
10. Denoising Autoencoder: This method uses deep learning models to impute missing values.

In general, I would recommend using multiple imputation or Expectation-Maximization (EM) as they are more robust methods that can account for the uncertainty in the imputed values, but it depends on the context, the amount of missing data, and the assumptions of the model. If the missing data is MCAR, then deletion or mean/mode imputation can be used, but if the missing data is not MCAR, then multiple imputation, predictive imputation, hot-deck imputation, EM imputation, regression imputation, interpolation or Denoising Autoencoder should be considered. Cold-deck imputation might be useful when the sample size is very small or there is a limited amount of data to work with.

12. What is A/B testing?

Answer: A/B testing, also known as split testing or bucket testing, is a method of comparing two versions of a product or service (A and B) to determine which one performs better. The goal of A/B testing is to determine which version of a product or service is more effective by randomly exposing different versions of the product or service to similar groups of users and measuring the response. The version that performs better is considered the "winner" and is usually implemented for the entire user base.

A/B testing is used in many fields, including web design, marketing, software development, and product management. It can be used to test website design, headlines, call-to-action buttons, and other elements of a website to see which version of a design or layout leads to more conversions or engagement. A/B testing can also be used to test different marketing campaigns to determine which one is more effective in driving sales or

conversions.
A/B testing is a statistical method that requires a sample size large enough to detect a statistically significant difference between two versions. The sample should be randomly and evenly divided between the two versions being tested. The results are then analyzed using statistical methods to determine which version performed better.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation of missing data is a simple and easy-to-implement technique, but it may not always be an acceptable practice, particularly if the missing data is not missing completely at random (MCAR). When missing data is not MCAR, imputing the mean can lead to biased estimates and inaccurate conclusions. Mean imputation assumes that the missing data is from the same distribution as the observed data, which is not always the case. If the missing data is not from the same distribution, the imputed mean would not represent the true mean of the population, which would lead to biased estimates of parameters. Additionally, mean imputation can lead to underestimation of variances and covariances, which can lead to invalid statistical inferences.
In general, if the amount of missing data is small or if the missing data is MCAR, then mean imputation can be an acceptable practice. However, if the amount of missing data is large or if the missing data is not MCAR, then more robust methods such as multiple imputation or Expectation-Maximization (EM) should be considered.

14. What is linear regression in statistics?

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the response variable or outcome variable) and one or more independent variables (also known as predictor variables or explanatory variables). The goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple independent variables) that describes the relationship between the dependent and independent variables.
In simple linear regression, the relationship between the dependent variable and the independent variable is modeled using a linear equation of the form:
$Y = a + bX$
Where Y is the dependent variable, X is the independent variable, a is the y-intercept (the point at which the line crosses the y-axis) and b is the slope of the line (the change in Y for a given change in X).
In multiple linear regression, the relationship between the dependent variable and multiple independent variables is modeled using a linear equation of the form:
$Y = a + b1X1 + b2X2 + ... + bnXn$
Where Y is the dependent variable, X1, X2, ... Xn are the independent variables, a is the y-intercept and b1, b2, ... bn are the regression coefficients that indicate the effect of each independent variable on the dependent variable.
Linear regression assumes linearity between the independent and dependent variables, which means that the relationship between the variables is linear. It also assumes

15. What are the various branches of statistics?

Answer: Statistics is a broad field that encompasses various branches that are used to solve different problems. Some of the main branches of statistics include:

1. Descriptive Statistics: This branch of statistics deals with summarizing, describing and presenting data. It includes techniques such as frequency distributions, measures of central tendency (mean, median, mode), measures of dispersion (variance, standard deviation, range), and graphical representations of data (histograms, bar charts, scatterplots, etc.).
2. Probability and Probability Distributions: This branch of statistics deals with the study of randomness and uncertainty. It includes concepts such as probability theory, random variables, probability distributions (normal, binomial, Poisson, etc.), and statistical inference.
3. Inferential Statistics: This branch of statistics deals with making inferences and predictions about a population based on a sample of data. It includes techniques such as estimation, hypothesis testing, and confidence intervals.
4. Regression Analysis: This branch of statistics deals with modeling the relationship between a dependent variable and one or more independent variables. It includes linear regression, multiple regression, logistic regression, and non-linear regression.
5. Time Series Analysis: This branch of statistics deals with analyzing time-based data. It includes

techniques for analyzing trends, seasonal patterns, and forecasting future values.

6. Multivariate Statistics: This branch of statistics deals with analyzing data with multiple variables. It includes techniques such as principal component analysis, factor analysis, cluster analysis, and multivariate regression.
7. Bayesian Statistics: This branch of statistics deals with the application of Bayesian probability theory to statistical inference. It includes techniques such as Bayesian estimation, Bayesian hypothesis testing, and Bayesian model selection.
8. Survival Analysis: This branch of statistics deals with analyzing time-to-event data. It includes techniques such as Kaplan-Meier estimation, Cox regression, and parametric survival models.

These are some of the main branches of statistics but not exhaustive, and there are other branches such as stochastic process, mathematical statistics, and so on.