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
   Ans. a) True

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
   Ans. a) Central Limit Theorem

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
   Ans. b) Modeling bounded count data

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
   Ans. 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
   Ans. c) Poission

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

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

Ans. a) 0

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

Ans. c) Outliers cannot conform to 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?

Ans. The normal distribution is a probability distribution that is widely used in statistical analysis and modeling. It is also known as the Gaussian distribution or the bell curve, because of its characteristic bell-shaped curve when plotted.

The normal distribution is characterized by two parameters: the mean (μ) and the standard deviation (σ). The mean represents the central tendency or average of the distribution, and the standard deviation represents the spread or variability of the distribution. The normal distribution is symmetric around the mean, and it has a bell-shaped curve that is smooth and continuous.

Many natural phenomena and statistical processes follow the normal distribution, such as measurement errors, physical and biological traits, test scores, and financial returns. The normal distribution is important in statistical inference, hypothesis testing, and estimation, because it has many desirable mathematical properties, such as the central limit theorem, which states that the sum or average of a large number of independent and identically distributed random variables tends to follow the normal distribution, regardless of the distribution of the individual variables.

The normal distribution is widely used in statistical analysis and modeling, because it allows for precise probabilistic calculations and predictions, and it provides a common framework for comparing and interpreting data across different domains and contexts.

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

Ans. Handling missing data is an important step in statistical analysis and modeling, because missing data can bias the results, reduce the sample size, and limit the generalizability of the findings. There are several approaches to handling missing data, and the choice of method depends on the nature and extent of the missingness, the research question, and the assumptions of the statistical model.

One common approach to handling missing data is to simply remove the cases with missing data, which is known as complete-case analysis or list-wise deletion. However, this method can lead to biased or inefficient estimates, especially if the missingness is related to the outcome or the predictors, or if the missing data are not missing completely at random (MCAR) but missing at random (MAR) or missing not at random (MNAR). Therefore, other imputation techniques are often recommended to handle missing data.

Imputation techniques involve replacing the missing values with estimated or imputed values, based on the available information and assumptions about the data. Some common imputation techniques include:

1. Mean or median imputation: This method involves replacing the missing values with the mean or median of the non-missing values of the variable. This method is simple and fast, but it assumes that the missing values have the same distribution as the non-missing values, which may not be true in all cases.
2. Regression imputation: This method involves using a regression model to predict the missing values, based on the available predictors and the non-missing values of the variable. This method can produce more accurate imputations than mean imputation, but it assumes that the regression model is correctly specified and that the missing values are MAR.
3. Multiple imputation: This method involves generating multiple imputed datasets, based on a probabilistic model of the missing data and the observed data. This method can handle missing data that are MAR or MNAR, and it produces unbiased and efficient estimates of the parameters of interest. However, this method requires more computational resources and expertise than other methods.

In general, the choice of imputation technique depends on the assumptions of the statistical model, the complexity of the missingness pattern, the amount of missing data, and the goals of the analysis. Multiple imputation is often recommended as a best practice, because it accounts for the uncertainty and variability of the imputed values and produces more robust and reliable results than other methods. However, any imputation method should be carefully evaluated and justified based on the assumptions and limitations of the data and the model.

12. What is A/B testing?

Ans. 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 or users to two groups, where one group receives the original or control version of the product or campaign, and the other group receives a modified or experimental version of the product or campaign.

The purpose of A/B testing is to measure and compare the effect of the two versions on a specific outcome or metric, such as conversion rate, click-through rate, engagement, or revenue. By comparing the performance of the two versions, A/B testing can provide insights into which design, feature, copy, or strategy is more effective, and can help businesses and organizations make data-driven decisions about how to optimize their products or services.

A/B testing typically involves several steps, including defining the hypothesis and the outcome metric, designing the control and experimental versions, randomly assigning participants or users to the two groups, collecting and analyzing the data, and drawing conclusions about the significance and the practical importance of the difference between the two versions. A/B testing can be conducted using various tools and platforms, such as Google Optimize, Optimizely, or VWO, and can be applied to different types of products or campaigns, such as websites, emails, ads, or landing pages.

13.Is mean imputation of missing data acceptable practice?

Ans. Mean imputation is a common method for handling missing data, but it has some limitations and potential drawbacks. Mean imputation involves replacing the missing values with the mean value of the non-missing values for that variable.

One of the main limitations of mean imputation is that it assumes that the missing values are missing at random (MAR), which means that the probability of a value being missing depends only on the observed values and not on the unobserved values. If the missing values are not MAR, then mean imputation can introduce bias and distort the distribution and the relationships between variables.

Moreover, mean imputation does not account for the uncertainty or variability in the missing values, and it can underestimate the standard errors and inflate the significance of the estimates. It can also lead to an underestimation of the variance and the correlations, and to a loss of power and precision in the analysis.

Therefore, mean imputation should be used with caution, and alternative methods such as multiple imputation, maximum likelihood estimation, or Bayesian approaches, may be more appropriate depending on the nature and the extent of the missing data, the type of analysis, and the research question.

14. What is linear regression in statistics?

Ans. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find a linear equation that best describes the relationship between the variables and can be used to predict the value of the dependent variable based on the values of the independent variables.

The linear equation takes the form of Y = a + bX, where Y is the dependent variable, X is the independent variable, a is the intercept (the value of Y when X is 0), and b is the slope (the change in Y per unit change in X). The slope represents the strength and direction of the relationship between the variables, and can be positive (when the variables increase together), negative (when one variable increases while the other decreases), or zero (when there is no relationship).

Linear regression can be used for both simple regression (when there is only one independent variable) and multiple regression (when there are two or more independent variables). The quality of the fit of the regression line can be measured by the coefficient of determination (R-squared), which represents the proportion of the variance in the dependent variable that is explained by the independent variables.

Linear regression can be used for various purposes, such as predicting sales, estimating the impact of a marketing campaign, analyzing the relationship between education and income, or identifying factors that affect customer satisfaction. It is widely used in business, economics, social sciences, and other fields.

15. What are the various branches of statistics?

Ans. Statistics can be broadly divided into two main branches: descriptive statistics and inferential statistics.

Descriptive statistics involves the collection, organization, summarization, and presentation of data. It includes measures such as mean, median, mode, standard deviation, range, and frequency distributions. Descriptive statistics is used to provide a summary of the data and to identify patterns, trends, and relationships among the variables.

Inferential statistics involves making generalizations about a population based on a sample of data. It includes techniques such as hypothesis testing, confidence intervals, and regression analysis. Inferential statistics is used to test hypotheses, make predictions, and draw conclusions about the population based on the sample data.

There are also various sub-branches of statistics, including:

1. Biostatistics: the application of statistical methods to analyze data related to medicine, public health, and biology.
2. Econometrics: the application of statistical methods to analyze economic data.
3. Psychometrics: the application of statistical methods to analyze psychological data, such as personality traits, intelligence, and attitudes.
4. Social statistics: the application of statistical methods to analyze social data, such as demographic, economic, and political data.
5. Statistical computing: the development and application of computer algorithms and software for statistical analysis.
6. Environmental statistics: the application of statistical methods to analyze environmental data, such as air and water quality, climate, and ecology.
7. Quality control and assurance: the use of statistical methods to monitor and improve the quality of products and services in industries such as manufacturing, healthcare, and education.