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

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

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data

4. Point out the correct statement.

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

5. ______ random variables are used to model rates.

c) Poisson

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

b) False

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

b) Hypothesis

8. Normalized data are centered at______and have units equal to standard deviations of the original data.

a) 0

9. Which of the following statement is incorrect with respect to outliers?

c) Outliers cannot conform to the regression relationship

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?

Normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is widely used in statistics and probability theory. The normal distribution is characterized by a bell-shaped curve, with the mean and the standard deviation determining its shape and location. The distribution is symmetric, with the highest point of the curve located at the mean of the distribution. The standard deviation of the distribution controls the spread or variability of the data.

Many natural phenomena and processes, such as the measurement of physical characteristics like height and weight, follow the normal distribution. The normal distribution has many important properties that make it useful in statistical analysis, including the ability to calculate probabilities and confidence intervals, the central limit theorem, and the ability to test hypotheses using z-scores. Many statistical methods, such as t-tests and ANOVA, assume that the underlying population follows a normal distribution.

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

Handling missing data is a common problem in statistical analysis and data science. There are several ways to handle missing data, including:

1. Complete case analysis: This involves simply removing any observations that have missing values. This method is simple and easy to implement, but can lead to biased results if the missing data are not missing at random.
2. Mean imputation: This involves replacing missing values with the mean value of the non-missing data for that variable. This method is easy to implement but can lead to biased results and underestimation of the variance.
3. Regression imputation: This involves using regression models to predict missing values based on other variables in the dataset. This method can produce accurate imputations but may require a significant amount of data and computational resources.
4. Multiple imputation: This involves creating multiple imputations for missing data, based on a statistical model that takes into account the uncertainty in the imputation process. Multiple imputation can produce more accurate estimates and account for the uncertainty in the imputed values.

The choice of imputation technique depends on the nature and extent of the missing data, the type of analysis being performed, and the assumptions underlying the imputation method. In general, multiple imputation is considered to be the most robust method for handling missing data, but it may also be the most computationally intensive. It is important to carefully evaluate the impact of missing data on the analysis results and choose an appropriate imputation method that minimizes bias and maximizes the accuracy of the analysis.

12. What is A/B testing?

A/B testing, also known as split testing, is a statistical method used in marketing, advertising, and product development to compare two versions of a product, advertisement, or web page to determine which one performs better in terms of a specific goal or outcome.

In A/B testing, two variants, A and B, are randomly assigned to different groups of users, customers, or participants, and their performance is measured and compared based on a specific metric, such as click-through rate, conversion rate, or sales. The groups should be chosen to be similar in their characteristics and size, so that the results are not biased by external factors.

The purpose of A/B testing is to determine whether one variant is more effective in achieving the desired outcome than the other, and to identify factors that may affect the outcome. A/B testing can be used to optimize various aspects of a product or marketing campaign, such as design, messaging, pricing, or layout.

A/B testing can be a powerful tool for businesses to improve their performance and achieve their goals. However, it is important to design the test carefully, control for confounding variables, and ensure that the sample size is large enough to detect meaningful differences between the variants.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is a simple method for handling missing data, where missing values are replaced by the mean of the observed values for that variable. While mean imputation is widely used and can be easy to implement, it has some limitations and may not always be an acceptable practice.

One of the main limitations of mean imputation is that it assumes that the missing values are missing completely at random (MCAR). If the missing values are not MCAR, and their presence or absence is related to other variables in the dataset, mean imputation can lead to biased estimates and distorted results.

Additionally, mean imputation can underestimate the variability of the data, since it does not account for the uncertainty introduced by imputing missing values. This can lead to an underestimation of standard errors and confidence intervals, and affect the results of subsequent analyses.

Despite these limitations, mean imputation can be acceptable in some situations, such as when the missing data are small in number and spread across the dataset, or when the missing data are believed to be MCAR. However, more sophisticated imputation methods, such as multiple imputation, may be more appropriate in situations where the missing data are substantial or not believed to be MCAR.

In summary, while mean imputation can be a useful method for handling missing data, its appropriateness depends on the assumptions underlying the imputation method and the nature of the missing data. Careful consideration of these factors is important to ensure that the results of the analysis are not biased or distorted by the missing data.

14. What is linear regression in statistics?

Linear regression is a statistical method used to model the relationship between two variables, where one variable is considered to be the dependent variable and the other variable is considered to be the independent variable. Linear regression assumes that there is a linear relationship between the variables, and aims to estimate the slope and intercept of the line that best fits the observed data.

In simple linear regression, there is only one independent variable and one dependent variable, and the goal is to find the best-fit line that describes the relationship between the variables. The equation of the line is represented as:

$y = \beta0 + \beta1x + \varepsilon$

where y is the dependent variable, x is the independent variable, β0 is the intercept, β1 is the slope, and ε is the error term.

The slope β1 represents the change in the dependent variable for a unit change in the independent variable, while the intercept β0 represents the value of the dependent variable when the independent variable is zero. The error term ε represents the unexplained variation in the dependent variable that is not accounted for by the linear relationship with the independent variable.

Linear regression can be used for both prediction and inference. In predictive regression, the goal is to use the model to make predictions about the dependent variable based on the values of the independent variable. In inferential regression, the goal is to use the model to make inferences about the relationship between the variables and to test hypotheses about the slope and intercept of the line.

15. What are the various branches of statistics?

Statistics is a broad field that includes a variety of sub-disciplines and branches. Some of the main branches of statistics include:

1. Descriptive statistics: This branch involves the collection, organization, and summarization of data using measures such as mean, median, mode, variance, and standard deviation.
2. Inferential statistics: This branch involves using data from a sample to make inferences or generalizations about a population. It involves methods such as hypothesis testing, confidence intervals, and regression analysis.
3. Biostatistics: This branch involves the application of statistical methods to analyze data related to health, medicine, and biology.
4. Business statistics: This branch involves the application of statistical methods to analyze data related to business and economics. It includes areas such as market research, financial analysis, and quality control.
5. Social statistics: This branch involves the application of statistical methods to analyze data related to social sciences such as sociology, political science, and psychology.
6. Bayesian statistics: This branch involves using Bayesian methods to analyze and interpret data. Bayesian statistics involves assigning probabilities to hypotheses or events, and updating these probabilities as new data becomes available.
7. Time series analysis: This branch involves the analysis of data that is collected over time, such as stock prices or weather patterns. Time series analysis involves methods such as trend analysis, seasonal analysis, and forecasting.
8. Multivariate statistics: This branch involves the analysis of data that involves more than two variables. It includes methods such as factor analysis, principal component analysis, and cluster analysis.

These are some of the main branches of statistics, but there are many other specialized sub-disciplines and branches, such as spatial statistics, ecological statistics, and computational statistics, among others.