STATISTICS WORKSHEET 1 ANSWERS :

Q1]

a) True

Q2]

a) Central Limit Theorem

Q3]

c) Modeling contingency tables

Q4]

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

Q5]

c) Poisson.

Q6]

b) False.

Q7]

b) Hypothesis

Q8]

c) 1.

Q9]

c) Outliers cannot conform to the regression relationship.

Q10]

Normal distribution, also known as Gaussian distribution or bell curve, is a continuous probability distribution that is symmetrical around the mean. It is characterized by its bell-shaped curve, which is defined by the mean and standard deviation of the data.

In a normal distribution, the mean, median, and mode are all equal, and the data points are clustered around the mean. The tails of the distribution extend indefinitely in both directions, and the area under the curve represents the total probability.

The normal distribution is widely used in statistics and probability theory because it describes many natural phenomena, such as the distribution of heights, weights, and IQ

scores. It is also used in hypothesis testing, confidence intervals, and other statistical analyses.

The normal distribution is defined by the probability density function (PDF), which is given by:

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

Q11]

There are several techniques for handling missing data in statistics and machine learning. Here are some common methods:

1. Deletion methods: These methods involve removing the observations with missing data. There are two types of deletion methods: listwise deletion and pairwise deletion. Listwise deletion removes any observation with missing data, while pairwise deletion removes only the missing data for each variable separately. However, deletion methods can lead to a loss of information and may introduce bias.
2. Imputation methods: These methods involve replacing missing data with estimated values. Here are some common imputation techniques:

a. Mean imputation: Replacing missing values with the mean of the non-missing values for that variable. This method is simple but can lead to biased estimates and underestimation of variance.

b. Median imputation: Replacing missing values with the median of the non-missing values for that variable. This method is more robust to outliers than mean imputation.

c. Mode imputation: Replacing missing values with the mode of the non-missing values for that variable. This method is used for categorical variables.

d. Regression imputation: Using regression analysis to predict missing values based on other variables in the dataset.

e. Multiple imputation: Creating multiple copies of the dataset and replacing missing values with different estimates in each copy. The results are then combined to obtain a final estimate.

f. Iterative imputation: Using machine learning algorithms to estimate missing values iteratively.

Q12]

A/B testing is a statistical method used to compare two versions of a product, website, or marketing campaign to determine which one performs better. It involves randomly assigning users to two groups: an experimental group that receives the new version (A) and a control group that receives the original version (B). By comparing the performance metrics of the two groups, such as conversion rates or click-through rates, A/B testing can help identify which version is more effective and inform data-driven decision making.

Here's an example of how A/B testing works:

1. Define the hypothesis: Start by defining a hypothesis that you want to test. For example, "Changing the color of the call-to-action button will increase the conversion rate."
2. Create two versions: Create two versions of the product, website, or marketing campaign. One version will be the original (control) and the other will be the new version (treatment).
3. Randomly assign users: Randomly assign users to either the control group or the treatment group. This ensures that any differences in performance are due to the treatment and not other factors.
4. Collect data: Collect data on the performance metrics for each group. This can include metrics such as conversion rates, click-through rates, or revenue.

5. Analyze the results: Use statistical analysis to compare the performance of the two groups. This can involve calculating the difference in means, confidence intervals, or p-values.
6. Make a decision: Based on the results of the analysis, make a decision about which version to use. If the treatment group performs significantly better than the control group, then the new version is likely to be more effective.

Q13]

Mean imputation is a simple and commonly used method for handling missing data, but it is not always an acceptable practice. Here are some factors to consider when deciding whether to use mean imputation:

1. Missing data mechanism: Mean imputation assumes that the missing data is missing completely at random (MCAR), which means that the missingness is unrelated to the values of the variables. If the missing data is missing at random (MAR) or missing not at random (MNAR), mean imputation can introduce bias and lead to incorrect conclusions.
2. Impact on variance: Mean imputation can lead to underestimation of variance, which can affect the accuracy of statistical inference. This is because the imputed values are assumed to be known without error, which is not the case.
3. Impact on correlation: Mean imputation can also affect the correlation between variables, especially if the missingness is related to the values of the variables. This can lead to incorrect conclusions about the relationships between variables.
4. Type of data: Mean imputation is generally not recommended for categorical data or data with skewed distributions. For categorical data, mode imputation may be more appropriate, while for skewed data, median imputation may be more appropriate.

Despite these limitations, mean imputation can be a useful method for handling missing data in some cases. For example, if the missing data is MCAR and the proportion of missing data is small, mean imputation may be acceptable. Additionally, mean imputation can be useful as a preliminary step to explore the data and identify patterns of missingness.

Q14]

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a type of supervised learning algorithm that is commonly used for regression tasks, where the goal is to predict a continuous outcome variable.

In linear regression, the relationship between the dependent variable and the independent variables is modeled as a linear function. Specifically, the dependent variable is assumed to be a linear combination of the independent variables, with an added error term. The coefficients of the independent variables are estimated using a method called least squares, which minimizes the sum of the squared errors between the predicted and actual values of the dependent variable.

Here's an example of a simple linear regression model with one independent variable:

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

In this equation, y is the dependent variable, x is the independent variable, β0 is the intercept term, β1 is the coefficient for the independent variable, and ε is the error term.

Linear regression can be extended to include multiple independent variables, in which case it is called multiple linear regression. The equation for multiple linear regression is:

$y = \beta0 + \beta1x1 + \beta2x2 + ... + \betanxn + \varepsilon$

In this equation, y is the dependent variable, x1, x2, ..., xn are the independent variables, β0 is the intercept term, β1, β2, ..., βn are the coefficients for the independent variables, and ε is the error term.

Linear regression is a powerful tool for modeling the relationship between variables and making predictions. However, it is important to carefully evaluate the assumptions of linear regression, such as linearity, independence, homoscedasticity, and normality, and

to consider alternative models if these assumptions are not met. Additionally, it is important to interpret the results of linear regression with caution and consider the potential limitations and biases.

Q15]

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

Descriptive statistics involves the collection, organization, and presentation of data in an informative way. It includes various measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and measures of shape (skewness, kurtosis). Descriptive statistics is used to summarize and describe the main features of a data set.

Inferential statistics, on the other hand, involves making inferences or predictions about a population based on a sample of data. It includes various statistical methods such as hypothesis testing, confidence intervals, regression analysis, and analysis of variance (ANOVA). Inferential statistics is used to make generalizations about a population based on a sample, and to test hypotheses or theories about the relationships between variables.

Within these two main branches, there are several subfields of statistics, including:

- Probability theory: This branch of statistics deals with the study of uncertainty and randomness. It includes various concepts such as probability distributions, random variables, and stochastic processes.
- Applied statistics: This branch of statistics deals with the application of statistical methods to real-world problems. It includes various fields such as biostatistics, econometrics, psychometrics, and sociometrics.

- Theoretical statistics: This branch of statistics deals with the development of statistical methods and theory. It includes various fields such as mathematical statistics, statistical learning, and statistical decision theory.
- Computational statistics: This branch of statistics deals with the development and application of computational methods for statistical analysis. It includes various fields such as machine learning, data mining, and statistical computing.