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

   Answer:(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

   Answer::(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

   Answer::(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

   Answer:(C) The square of a standard normal random variable follows what is called chi-squareddistribution

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

   Answer::(C) poisson

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

   Answer::(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

   Answer::(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

   Answer::(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

- **Answer::(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?

The normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is symmetric and follows a specific mathematical form. It is a fundamental concept in statistics and probability theory and has widespread applications in various fields.

The characteristics of a normal distribution are as follows:

1. **Symmetry:** The normal distribution is symmetric, which means it is evenly distributed around its mean The mean, median, and mode of a normal distribution are all equal and located at the center of the distribution.

2. **Bell-shaped curve:** The probability density function (PDF) of a normal distribution has a distinctive bell-shaped curve. The curve is unimodal, meaning it has a single peak. The shape of the curve is determined by its mean and standard deviation.

3. **Mean and standard deviation:** The mean (μ) specifies the center of the distribution, while the standard deviation (σ) determines the spread or dispersion of the data around the mean. The variance of a normal distribution is the square of the standard deviation.

4. **Empirical Rule:** The normal distribution follows the empirical rule (also known as the 68-95-99.7 rule), which states that approximately 68% of the data falls within one standard deviation of the mean, about 95% falls within two standard deviations, and nearly 99.7% falls within three standard deviations.

The normal distribution is widely used in statistical inference, hypothesis testing, and modeling in various fields such as economics, finance, social sciences, and natural sciences. Many real-world phenomena, such as the height and weight of individuals, errors in measurements, and IQ scores, tend to exhibit a normal distribution.The properties and mathematical simplicity of the normal distribution make it a valuable tool in statistical analysis, providing a basis for understanding and describing random variables and their distributions.

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

Handling missing data is an important aspect of data preprocessing and analysis. There are several techniques for handling missing data, and the choice of technique depends on the nature of the data and the specific analysis objectives. some commonly used techniques for handling missing data:

1. **Deletion:** This approach involves removing the samples or variables with missing data.

2. **Mean or Median Imputation:** This technique involves replacing missing values with the mean or median of the available data for that variable. It is a simple approach but may underestimate the variability in the data and introduce bias.

3. **Mode Imputation:** Mode imputation is used for handling missing values in categorical variables. It involves replacing missing values with the mode (most frequent value) of the available data for that variable.

The choice of imputation technique depends on various factors, including the missing data mechanism, the distribution of the data, the amount of missingness, and the specific analysis requirements

12. What is A/B testing?

A/B testing, also known as split testing or bucket testing, is a method used to compare two versions of a webpage, app, or marketing campaign to determine which one performs better in terms of predefined metrics or goals. It is a statistical hypothesis testing technique that helps businesses make data-driven decisions and optimize their strategies.

Here's how A/B testing typically works:

1. **Two Versions:** Two versions, often referred to as the control and the variant, are created with a single differing element. This element could be a design change, layout modification, content variation, pricing difference, or any other specific feature.

2. **Random Assignment:** Users or participants are randomly divided into two groups: one group experiences the control version (A), and the other group experiences the variant version (B). Random assignment helps ensure that any differences observed between the groups are due to the version they were exposed to and not other factors.

3. **Data Collection:** Relevant data and user interactions are collected for each group, such as click-through rates, conversion rates, bounce rates, or any other metric that measures the desired outcome. The data is typically tracked using analytics tools or custom tracking codes.

4. **Statistical Analysis:** Statistical analysis is performed on the collected data to determine if there is a statistically significant difference between the control and variant versions. Commonly used statistical techniques include hypothesis testing, t-tests, chi-square tests, or regression analysis.

5. **Decision Making:** Based on the analysis results, conclusions are drawn regarding which version performed better. The version with a statistically significant improvement in the desired metrics or goals is usually considered the winner and is implemented or further optimized.

A/B testing allows businesses to make data-informed decisions by providing evidence of how changes impact user behavior and key performance indicators. It helps to validate assumptions, improve user experience, increase conversions, optimize marketing campaigns, and ultimately drive business growth.

It's important to note that A/B testing requires careful planning, proper sample sizes, statistical rigor, and adherence to ethical considerations. Additionally, it is crucial to focus on testing one variable at a time to isolate the impact of that particular change.

13. Is mean imputation of missing data acceptable practice?

Mean imputation, where missing values are replaced with the mean of the available data, is a simple and commonly used method for handling missing data. However, whether mean imputation is an acceptable practice depends on several factors and should be used with caution. Here are some considerations:

1. **Potential Bias:** Mean imputation can introduce bias in the data, especially when the missing data are not missing completely at random (MCAR). If there is a systematic relationship between the missing values and other variables, mean imputation may distort the distribution and relationships within the data.

2. **Underestimation of Variability:** Mean imputation does not account for the uncertainty associated with the missing values. It assumes that the imputed values are known with certainty and ignores the variability of the missing data. This can lead to an underestimation of the true variability in the data, affecting subsequent statistical analysis.

3. **Impact on Relationships:** Mean imputation can artificially strengthen or weaken relationships between variables. By imputing missing values with the mean, the imputed values tend to be closer to the mean, potentially reducing the observed variability and impacting correlation or regression analyses.

Despite these limitations, mean imputation may still be used in certain scenarios, such as when the missingness is assumed to be missing completely at random (MCAR), and the proportion of missing values is relatively small compared to the total dataset.

It is generally recommended to explore more sophisticated imputation techniques, such as multiple imputation or regression imputation, which can provide more accurate results by considering the relationships between variables and accounting for uncertainty in the imputed values. These techniques can yield better results and produce more valid inferences compared to simple mean imputation.

14. What is linear regression in statistics?

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fit linear equation that describes the linear relationship between the variables.

In linear regression, the dependent variable is the variable of interest or the outcome that we want to predict or explain. The independent variables, also known as predictor variables or features, are used to explain or predict the value of the dependent variable.

The linear regression model assumes a linear relationship between the dependent variable and the independent variables, which can be represented by a straight line in a two-dimensional space or a hyperplane in higher-dimensional spaces.

The general equation for a simple linear regression, involving a single independent variable, can be expressed as:

$y = \beta_0 + \beta_1 x_1 + \varepsilon$

In this equation:
- y represents the dependent variable.
- x_1 represents the independent variable.
- β_0 and β_1 are the coefficients or parameters estimated by the regression model. β_0 is the intercept, which represents the value of y when x_1 is zero. β_1 is the slope, indicating the change in y for a unit change in x_1.
- ε represents the error term or residual, which captures the unexplained variation in the dependent variable that is not accounted for by the linear relationship with the independent variable.

The goal of linear regression is to estimate the coefficients (β_0 and β_1) that minimize the sum of squared residuals, also known as the least squares method. This estimation is typically done using optimization techniques or closed-form solutions such as the Normal Equation.

Linear regression is widely used in various fields for prediction, forecasting, and understanding the relationship between variables. It provides valuable insights into the direction, strength, and significance of the relationship, and allows for making predictions or estimating the value of the dependent variable based on the values of the independent variables.

15. What are the various branches of statistics?

Statistics is a broad field with various branches or subfields that focus on different aspects of data analysis, inference, and application. Some of the main branches of statistics include:

1. **Descriptive Statistics:** Descriptive statistics involves summarizing and describing data using measures such as central tendency (mean, median, mode) and variability (standard deviation, range). It provides techniques for organizing, presenting, and describing data sets.

2. **Inferential Statistics:** Inferential statistics deals with drawing conclusions and making inferences about populations based on sample data. It includes hypothesis testing, confidence intervals, and estimation techniques. Inferential statistics allows us to make generalizations and infer population characteristics from sample data.

3. **Probability Theory:** Probability theory is the foundation of statistics and deals with the study of uncertainty and randomness. It includes concepts such as probability distributions, random variables, and probability calculations. Probability theory is used to model and analyze the uncertainty associated with events and outcomes.

4. **Statistical Modeling:** Statistical modeling involves building mathematical models that represent the relationships between variables and data. It includes regression analysis, time series analysis, multivariate analysis, and other modeling techniques. Statistical models help understand and explain complex data patterns and make predictions or forecasts.

5. **Experimental Design:** Experimental design focuses on the planning and design of experiments to collect data and analyze the effects of different factors or treatments. It includes techniques such as randomization, control groups, factorial designs, and response surface methodology. Experimental design ensures the validity and efficiency of experiments and helps identify causal relationships.

6. **Bayesian Statistics:** Bayesian statistics is an approach to statistical inference that incorporates prior knowledge or beliefs about the data. It uses Bayes' theorem to update the prior beliefs based on observed data and calculate posterior probabilities. Bayesian statistics provides a framework for incorporating prior information and uncertainty into statistical analysis.

These are some of the main branches of statistics, and there are additional specialized areas such as spatial statistics, environmental statistics, financial statistics, social statistics, and more. Each branch has its own techniques, methodologies, and applications, but they are all interconnected and contribute to the broader field of statistics.