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

Correct 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

Correct 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

Correct answer: b) Modeling bounded count data

STATISTICS WORKSHEET-1

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

Correct answer: 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

Correct answer: c) Poisson

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

a) True

b) False

Correct answer: b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

STATISTICS WORKSHEET-1

Correct answer: b) Hypothesis

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

a) 0

b) 5

c) 1

d) 10

Correct 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

Correct 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 term "Normal Distribution," also known as the Gaussian distribution or bell curve, is a fundamental concept in statistics and probability theory. It describes a specific type of probability distribution for a continuous random variable, characterized by its symmetric bell-shaped curve. In a normal distribution:

1. The mean (average), median, and mode are all equal and located at the center of the distribution.
2. The curve is symmetric around the mean, meaning that the values on both sides of the mean are equally distributed.
3. The standard deviation determines the spread or variability of the data. A larger standard deviation leads to a wider curve, and a smaller standard deviation leads to a narrower curve.

STATISTICS WORKSHEET-1

The normal distribution is important because many natural phenomena and random variables tend to follow this distribution pattern. It's a foundation for many statistical methods and hypothesis tests.

Example: Height of Adults
A real-time example of a normal distribution is the distribution of heights of adult humans. In a large population, if you were to measure the heights of a significant number of adults and create a histogram of those heights, you would likely observe a bell-shaped curve. The majority of people would be clustered around the average height (mean), with fewer individuals being either significantly shorter or taller. This is a classic example of a normal distribution in the real world.

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

Handling missing data is an important aspect of data analysis and modeling, as incomplete data can lead to biased or inaccurate results. There are several techniques for handling missing data, and the choice of technique depends on the nature of the data and the analysis being conducted. Here are some common approaches to handling missing data, including imputation techniques:

Deletion Methods:
a. Listwise Deletion: Rows with missing values are removed entirely from the dataset. This can lead to loss of valuable information and potentially biased results, especially if the missing data is not completely random.
b. Pairwise Deletion: Only specific columns or variables with missing values are excluded from calculations, but the rest of the data is used. This approach retains more data but can lead to incomplete analyses.

Imputation Techniques:
a. Mean, Median, or Mode Imputation: Replace missing values with the mean, median, or mode of the observed values in the variable. This is a simple method but may not accurately represent the underlying distribution.
b. Predictive Modeling Imputation: Use other variables to predict and fill in missing values. Techniques such as linear regression, decision trees, or k-nearest neighbors can be used.
c. Interpolation and Extrapolation: For time series data, missing values can be estimated using interpolation (within observed time points) or extrapolation (beyond observed time points).
d. Multiple Imputation: Generate multiple plausible values for missing data based on the observed data's distribution. This method takes into account the uncertainty associated with missing data.
e. Hot-Deck Imputation: Assign missing values from similar observed values, often using some measure of similarity or distance.
f. Mean Substitution: Replace missing values with the mean of non-missing values within a defined group or category.

The choice of imputation technique depends on factors such as the type of data, the extent of missingness, the presence of patterns in missing data, and the goals of the analysis. It's important to be cautious when imputing data, as the chosen method can impact the validity of your results.

STATISTICS WORKSHEET-1

Always consider the potential biases and limitations introduced by imputing missing data, and report your imputation methods and any associated uncertainties transparently in your analysis. In addition, sensitivity analyses can help assess the robustness of your results to different imputation strategies.

12. What is A/B testing?

A/B testing, also known as split testing, is a method used in statistics and experimentation to compare two different versions of something (such as a webpage, advertisement, or product) and determine which one performs better in terms of a specific outcome. The goal of A/B testing is to make data-driven decisions by assessing the impact of changes and improvements.

Here's how A/B testing typically works:

Two Versions: You have two versions of something that you want to compare. One version is called the "control" (A), and the other is the "variant" (B), which includes some changes or modifications.

Randomization: You randomly divide your audience or sample into two groups: one group sees the control version, and the other group sees the variant version. This randomization helps ensure that the groups are comparable and that any differences observed are likely due to the changes in the variant.

Testing: Both versions are presented to their respective groups, and data is collected on a specific metric or outcome of interest. This could be anything from click-through rates, conversion rates, revenue, user engagement, or any other relevant performance metric.

Comparison: After a sufficient amount of data is collected (to achieve statistical significance), you analyze the results. You compare the performance of the control group (A) with that of the variant group (B) to determine which version performed better.

Conclusion: Based on the analysis, you can decide whether the changes made in the variant version led to a statistically significant improvement in the desired outcome. If the results are clear, you may implement the changes from the variant version.

A/B testing is widely used in various fields, including marketing, web design, product development, and user experience optimization. It allows organizations to make informed decisions about design, content, and features by quantifying the impact of changes on user behavior and business objectives. A well-executed A/B test helps minimize biases and provides valuable insights into how different variations affect user behavior and preferences.

STATISTICS WORKSHEET-1

13. Is mean imputation of missing data acceptable practice?

Mean imputation of missing data is a simple and straightforward method, but its acceptability depends on the context and the nature of the data. While mean imputation has its advantages, it also comes with limitations and potential drawbacks that should be carefully considered before applying it:

Advantages of Mean Imputation:

Simplicity: Mean imputation is easy to implement and understand, making it a quick solution for handling missing data.
Preserves Sample Size: Mean imputation allows you to retain all cases in your analysis, as no data points are removed.
Limitations and Drawbacks:

Distortion of Variability: Mean imputation can underestimate the true variability of the data. It tends to artificially reduce the variance, potentially affecting statistical analyses and hypothesis testing.
Introduction of Bias: Mean imputation assumes that the missing values are missing completely at random (MCAR), which may not hold true in all cases. If the missingness is related to the variable being imputed or other variables, mean imputation can introduce bias into the analysis.
Altered Relationships: Imputing missing data with the mean can alter the relationships between variables, potentially leading to inaccurate results and conclusions.
Misleading Interpretation: Mean imputation can lead to misleading interpretations, especially if the proportion of missing data is substantial or if there are patterns in the missingness that are not addressed.
Given these limitations, mean imputation is generally not recommended for datasets with substantial missing data or when the missingness is related to the variable being imputed. In such cases, more advanced imputation methods that capture the underlying relationships and distribution of the data may be more appropriate.

If you choose to use mean imputation, it's crucial to be transparent about the method used in your analysis and to consider performing sensitivity analyses to assess the impact of imputation on your results. In cases where data quality and accuracy are critical, exploring alternative imputation methods such as predictive modeling imputation, multiple imputation, or domain-specific techniques may provide more reliable results.

14. What is linear regression in statistics?

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (also called the response or outcome variable) and one or more independent variables (also called predictor or explanatory variables). It assumes that there is a linear relationship between the variables, meaning that a change in the independent variable(s) is associated with a proportional change in the dependent variable.

The goal of linear regression is to find the best-fitting linear equation (a straight line) that describes the relationship between the variables. This equation can then be used for making predictions and understanding how changes in the independent variable(s) affect the dependent variable. The linear regression equation has the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

STATISTICS WORKSHEET-1

Where:
y is the dependent variable.
x_1, x_2, ..., x_k are the independent variables.
β_0, β_1, β_2, ..., β_k are the coefficients (parameters) that represent the effect of each independent variable on the dependent variable.
ε is the error term, representing the difference between the observed and predicted values.
The coefficients (β_0, β_1, β_2, etc.) are estimated from the data using a method that minimizes the sum of squared differences between the observed values and the values predicted by the linear equation. This method is often referred to as the "least squares" approach.

Linear regression can be categorized into two main types:

Simple Linear Regression: Involves only one independent variable. The equation takes the form: $y = \beta_0 + \beta_1 x + \varepsilon$.

Multiple Linear Regression: Involves two or more independent variables. The equation becomes a linear combination of multiple variables: $\mathbf{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon.}$

Linear regression is widely used in various fields such as economics, social sciences, engineering, and natural sciences for purposes like predicting future outcomes, understanding relationships between variables, and making data-driven decisions. It's important to note that while linear regression assumes a linear relationship between variables, it might not always be the most appropriate model for data with more complex relationships. In such cases, other regression techniques or machine learning methods may be more suitable.

15. What are the various branches of statistics?

In the field of data science, statistics plays a crucial role in extracting insights from data and making informed decisions. Here are some key branches of statistics as applied to data science:

Descriptive Statistics: Descriptive statistics involve summarizing and describing data using measures like mean, median, mode, standard deviation, percentiles, and graphical representations. They provide an initial understanding of the dataset's characteristics.

Inferential Statistics: Inferential statistics help draw conclusions and make predictions about a population based on a sample. Techniques include hypothesis testing, confidence intervals, and regression analysis.

Probability Theory: Probability concepts underpin many data science methods. Understanding probability distributions, conditional probability, Bayes' theorem, and random variables is crucial for modeling uncertainty.

Statistical Learning: Statistical learning includes techniques for building models to predict outcomes or understand relationships in data. This encompasses linear and logistic regression, decision trees, random forests, support vector machines, and more.

Bayesian Statistics: Bayesian methods involve updating beliefs or probabilities based on new data or evidence. Bayesian inference and modeling are used for various tasks in data science, including parameter estimation and uncertainty quantification.

STATISTICS WORKSHEET-1

Time Series Analysis: Time series methods are used to analyze data collected over time, such as stock prices, sales data, and sensor readings. Techniques include ARIMA models, exponential smoothing, and state space models.

Spatial Statistics: Spatial statistics deal with data that has a geographic component. It's applied in fields such as geospatial analysis, epidemiology, and urban planning to analyze spatial patterns and relationships.

Multivariate Analysis: Multivariate analysis explores relationships among multiple variables simultaneously. Techniques include principal component analysis (PCA), factor analysis, and canonical correlation analysis.

Experimental Design: Experimental design involves planning and conducting experiments to gather data in a controlled manner. Proper design ensures reliable and interpretable results.

Sampling Techniques: Sampling methods are used to select representative subsets of data from larger populations. Proper sampling is essential for making generalizations from data.

Statistical Computing: Proficiency in programming languages like R or Python is crucial for implementing statistical analyses, creating visualizations, and building models.

Data Preprocessing and Cleaning: Data preprocessing involves handling missing values, outliers, and noise. Cleaning ensures data quality before analysis.

Statistical Software: Proficiency in statistical software packages and libraries (e.g., R, Python's pandas and numpy) is necessary for data manipulation, analysis, and visualization.

Causal Inference: Causal inference methods aim to establish cause-and-effect relationships between variables. They're used to draw meaningful insights from observational data.

Machine Learning Interpretability and Fairness: Statistical techniques are essential for interpreting and explaining machine learning models, as well as for ensuring fairness and mitigating bias.

These branches of statistics intersect with various data science methodologies and techniques, allowing data scientists to extract meaningful insights, build predictive models, and drive data-driven decision-making in diverse domains.