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

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

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

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

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

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

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

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

Answers for Q1 to Q9:-

1. Bernoulli random variables take (only) the values 1 and
a) True
b) False
Answer: a

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

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

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

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

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

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

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

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

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:

The term "normal distribution," also known as the Gaussian distribution or the bell curve, refers to a specific type of probability distribution in statistics. It's a fundamental concept in many fields, including science, engineering, finance, and even psychology.

Key characteristics:

Bell-shaped curve: The most recognizable feature of a normal distribution is its visual representation, which resembles a bell. This curve depicts the probability of different values occurring within a dataset.

Symmetry: The curve is symmetrical around its mean, which is the average value of the data. This means that values above and below the mean are equally likely, but with decreasing probability as they get further away.

Concentration around the mean: Most of the data points in a normal distribution cluster around the mean. Specifically, about 68% of the data falls within 1 standard deviation of the mean, 95% within 2 standard deviations, and 99.7% within 3 standard deviations.

Standard deviation: This measure defines the spread of the curve. A higher standard deviation indicates a wider spread, with data points more likely to be further away from the mean.

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

Understanding Missing Data:

Missing Completely at Random (MCAR): Missingness has no relationship with other variables or the missing value itself.
Missing at Random (MAR): Missingness is related to other observed variables, but not the missing value itself.
Missing Not at Random (MNAR): Missingness is related to the missing value itself, often creating bias.

Common Strategies for Handling Missing Data:

1.Deletion:

Listwise Deletion: Removes entire rows with missing values. Simple but can lead to significant data loss.
Pairwise Deletion: Omits rows only for calculations involving variables with missing values. Preserves more data but can create analysis inconsistencies.

2.Imputation:

Mean/Median/Mode Imputation: Replaces missing values with the mean, median, or mode of the variable. Simple, but can distort distributions and relationships.
Random Sample Imputation: Replaces missing values with randomly selected observed values from the variable. Preserves variability but can introduce noise.
Predictive Imputation: Uses regression or machine learning to predict missing values based on other variables. More sophisticated, but requires careful model selection and validation.
Multiple Imputation (MI): Creates multiple plausible datasets with imputed values, accounts for uncertainty, often considered the most robust approach.

3.Other Techniques:

Interpolation (for time-series data): Predicts missing values based on patterns in observed data.
KNN Imputation: Uses nearest neighbors to impute missing values based on similar observations.
Matrix Completion: Reconstructs missing values in matrices using low-rank approximations.
Handling Missingness as a Feature: Creates a new feature indicating missingness, can capture patterns in missing data.

12.What is A/B testing?

A/B Testing:

A/B testing, also known as split testing or bucket testing, is a scientific method for comparing two versions of a variable, such as a web page, app feature, email, or ad, to see which one performs better. It's essentially an experiment where you show different variations of something to different groups of people and then measure which variation is more successful based on a predetermined set of criteria

Here's how it works:

Define your hypothesis: What do you think will happen if you change something on your webpage or app? For example, you might think that changing the color of your call-to-action button will increase clicks.
Create two versions of your webpage or app: One version will be the original (the control), and the other will have the change you want to test (the variation).

Split your traffic: Send some of your website or app visitors to the control version and some to the variation version.
Track your results: Use a tool like Google Analytics to track how users interact with each version of your webpage or app. You might track things like clicks, conversions, and bounce rate.
Analyze your results: Once you have enough data, you can use statistical analysis to see if there is a significant difference in performance between the two versions. If there is, you can then decide whether to implement the change on your live website or app.

13. Is mean imputation of missing data acceptable practice?

Mean Imputation:

Mean imputation, where missing values are replaced with the mean of the observed values, is a common method for handling missing data. While it is a simple and quick approach, it has both advantages and disadvantages, and its acceptability depends on the context and the nature of the data.

14.What is linear regression in statistics?

Linear Regression:

Linear regression is a statistical method used to model and analyze the linear relationship between two variables: a dependent variable (the outcome you want to predict) and one or more independent variables (also known as predictors or explanatory variables).

Types of linear regression:

Simple linear regression: One independent variable.
Multiple linear regression: Two or more independent variables.

Equation for simple linear regression (one independent variable):

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

where:-

β0 (intercept): The value of y when x = 0 (where the line crosses the y-axis).

β1 (slope): The change in y for every one-unit change in x (the steepness of the line).

ε (error term): Represents the difference between the actual data points and the predicted values on the line.

How it works:

Collect data: Gather paired observations of the dependent and independent variables.

Plot the data: Create a scatter plot to visualize the relationship between the variables.

Fit the regression line: Use statistical techniques (like least squares) to find the line that best fits the data, minimizing the distance between the line and the data points.

Interpret the results: Analyze the intercept, slope, and other statistical measures to understand the strength and significance of the relationship

15. What are the various branches of statistics?

In data science, statistics plays a crucial role in extracting meaningful insights from data. Various branches of statistics are particularly relevant to data science applications. Some of the key branches include:

Descriptive Statistics: Descriptive statistics involves methods for summarizing and presenting data. This includes measures such as mean, median, mode, range, variance, and standard deviation. Descriptive statistics are often the first step in understanding the characteristics of a dataset.

Inferential Statistics: Inferential statistics is used to make predictions or inferences about a population based on a sample of data. Techniques such as hypothesis testing, confidence intervals, and regression analysis fall under this category.

Probability Theory: Probability is fundamental to understanding uncertainty and randomness in data. Data scientists use probability theory to model and analyze the likelihood of different outcomes.

Bayesian Statistics: Bayesian statistics is an approach that involves updating probabilities based on prior knowledge and new evidence. It is increasingly used in data science for decision-making and updating beliefs in the face of new data.

Machine Learning Statistics: Many techniques in machine learning have statistical foundations. This includes supervised learning methods (e.g., linear regression, logistic regression), unsupervised learning methods (e.g., clustering), and probabilistic models.

Statistical Learning: Statistical learning is a broader field that includes machine learning but also encompasses traditional statistical methods. It focuses on understanding the principles and assumptions behind various learning algorithms.

Time Series Analysis: Time series analysis deals with data collected over time. It includes methods for analyzing trends, seasonality, and making predictions based on historical data.

Spatial Statistics: Spatial statistics involves the analysis of data with a spatial component, such as geographical locations. It is relevant in applications like geographic information systems (GIS) and spatial data analysis.

Resampling Methods: Resampling methods, such as bootstrapping and cross-validation, are used to assess the stability and reliability of statistical estimates. They are particularly valuable in the evaluation of predictive models.

Experimental Design: Experimental design is crucial for designing studies and experiments that yield reliable and valid results. It includes techniques for randomization, control group selection, and minimizing biases.