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, properlynormalized, 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-normaldistribution
   b) Sums of normally distributed random variables are again normally distributed even if the variablesare dependent
   c) The square of a standard normal random variable follows what is called chi-squareddistribution
   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 theoriginal 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

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, is a fundamental concept in statistics and probability theory. It describes a symmetric, bell-shaped probability distribution that is characterized by two parameters: the mean (μ) and the standard deviation (σ).

In a normal distribution:

The distribution is symmetric around the mean μ.

The mean, median, and mode of the distribution are all equal and located at the center of the distribution.

The standard deviation σ determines the spread or dispersion of the data points around the mean. A larger standard deviation implies a wider distribution, while a smaller standard deviation implies a narrower distribution.

The distribution is continuous and extends indefinitely in both directions.

The famous empirical rule (or 68-95-99.7 rule) states that approximately 68% of the data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations.

The normal distribution is widely used in various fields such as statistics, finance, natural sciences, social sciences, and engineering due to its mathematical tractability and its occurrence in many natural phenomena. Additionally, many statistical methods and hypothesis tests are based on the assumption of normality or approximate normality of data.

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

Handling missing data is a crucial step in data analysis to ensure that the results are reliable and accurate. There are several approaches to handling missing data, including:

Deletion: This involves removing observations with missing data. It can be done in two ways:

Listwise deletion (complete case analysis): Removing entire observations with any missing values.
Pairwise deletion: Analyzing only the available data for each variable pair.
Imputation: Imputation involves replacing missing values with estimated values. Some commonly used imputation techniques include:

a. Mean/Median/Mode Imputation: Replace missing values with the mean, median, or mode of the observed data for that variable.

b. Hot-Deck Imputation: Replace missing values with values from similar observed cases (e.g., nearest neighbor imputation).

c. Cold-Deck Imputation: Replace missing values with values from a different dataset.

d. Regression Imputation: Predict missing values using a regression model based on other variables.

e. Multiple Imputation: Generate multiple plausible values for each missing value, creating several complete datasets. Analyze each dataset separately, then combine the results.

f. K-Nearest Neighbors (KNN) Imputation: Replace missing values with the average of the K nearest

neighbors.

Advanced Techniques: Some advanced techniques involve using machine learning algorithms to predict missing values, such as decision trees or random forests.

12. What is A/B testing?

A/B testing, also known as split testing, is a method used to compare two versions of a webpage, app, email, or other digital asset to determine which one performs better. It is a randomized experiment with two variants, A and B, where variant A is the control and variant B is the treatment.

13. Is mean imputation of missing data acceptable practice?

Mean imputation, where missing values are replaced with the mean of the observed data for that variable, is a simple and commonly used technique for handling missing data. However, whether it is an acceptable practice depends on several factors and considerations:

Assumption of Missing Completely at Random (MCAR): Mean imputation assumes that the missing data are missing completely at random, meaning that the probability of data being missing is unrelated to both observed and unobserved data. If this assumption holds, mean imputation can provide unbiased estimates. However, if data are missing systematically (i.e., not at random), mean imputation can introduce bias into the analysis.

Impact on Variability: Mean imputation tends to underestimate the variability of the data because it artificially reduces variance by filling in missing values with the same value (the mean). This can affect the accuracy of statistical inference and lead to underestimation of standard errors and inflated Type I error rates.

Loss of Information: Mean imputation replaces missing values with a single value, ignoring any variability or patterns in the missing data. This can lead to loss of information and potentially distort the relationships between variables in the dataset.

Applicability to Different Types of Data: Mean imputation is more suitable for continuous variables with approximately symmetric distributions. For categorical variables or variables with skewed distributions, other imputation methods may be more appropriate.

Impact on Results: The choice of imputation method can influence the results of the analysis, including parameter estimates, hypothesis tests, and predictive models. It's essential to consider the potential impact of mean imputation on the validity and interpretation of the results.

14. What is linear regression in statistics?

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It is one of the simplest and most commonly used techniques for predictive modeling and understanding the relationship between variables.

In linear regression, the relationship between the dependent variable (often denoted as y)and the independent variable(s) (often denoted as x) is modeled using a linear equation of the form:

$y=\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn+\epsilon$

Where:

- yy is the dependent variable (the variable we want to predict or explain).
- $x1,x2,...,xnx1,x2,...,xn$ are the independent variables (also known as predictor variables or features).
- $\beta 0\beta 0$ is the intercept term (the value of yy when all xx's are zero).
- $\beta 1,\beta 2,...,\beta n\beta 1,\beta 2,...,\beta n$ are the coefficients (parameters) representing the change in yy for a one-unit change in each xx, holding other variables constant.
- $\epsilon\epsilon$ is the error term (residuals), representing the difference between the observed yy and the predicted yy by the model.

  The goal of linear regression is to estimate the coefficients (parameters) $\beta 0,\beta 1,...,\beta n\beta 0,\beta 1,...,\beta n$ that minimize the sum of squared differences between the observed values of the dependent variable and the values predicted by the linear equation. This is typically done using the method of least squares.

15. What are the various branches of statistics?

Statistics is a broad field that encompasses various branches, each focusing on different aspects of data collection, analysis, interpretation, and inference. Some of the main branches of statistics include:

1. **Descriptive Statistics**: Descriptive statistics involves methods for summarizing and describing the main features of a dataset. This includes measures of central tendency (e.g., mean, median, mode), measures of dispersion (e.g., variance, standard deviation), and graphical representations (e.g., histograms, box plots).
2. **Inferential Statistics**: Inferential statistics involves making inferences and drawing conclusions about populations based on sample data. This includes hypothesis testing, confidence intervals, and estimation techniques.
3. **Probability Theory**: Probability theory is the mathematical foundation of statistics, dealing with the study of random phenomena and uncertainty. It includes concepts such as probability distributions, random variables, independence, and conditional probability.
4. **Biostatistics**: Biostatistics focuses on the application of statistical methods to biological and health-related data. It plays a crucial role in clinical trials, epidemiological studies, genetics research, and public health.
5. **Econometrics**: Econometrics applies statistical methods to economic data to analyze economic relationships, test economic theories, and make forecasts. It is widely used in areas such as finance, macroeconomics, and microeconomics.
6. **Psychometrics**: Psychometrics is the branch of statistics concerned with the measurement of psychological attributes such as intelligence, personality, and attitudes. It involves the development and validation of measurement instruments (e.g., tests, surveys) and the analysis of psychological data.
7. **Spatial Statistics**: Spatial statistics deals with the analysis of data that have spatial or geographic attributes. It includes techniques for spatial data visualization, spatial autocorrelation, spatial interpolation, and spatial regression.
8. **Time Series Analysis**: Time series analysis focuses on the analysis of data collected over time. It includes methods for modeling and forecasting time-dependent processes, such as trends, seasonality, and autocorrelation.
9. **Multivariate Analysis**: Multivariate analysis deals with the analysis of datasets with multiple variables. It includes techniques for exploring relationships among variables, such as principal component analysis, factor analysis, and cluster analysis.
10. **Statistical Learning**: Statistical learning, also known as machine learning, involves the development of algorithms and techniques for building predictive models from data. It includes supervised learning (e.g., regression, classification) and unsupervised learning (e.g., clustering, dimensionality reduction).