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-squared distribution

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

   Answer: c) Poisson

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

   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

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

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?

Answer: The normal distribution, also known as the Gaussian distribution or bell curve, is a fundamental and widely used probability distribution in statistics. It's characterized by its symmetrical shape, where the majority of the data is concentrated near the mean value, and the distribution tapers off as you move away from the mean in both directions.

- **Symmetry:** The distribution is symmetric around the mean, meaning that the left and right sides of the mean are mirror images of each other.

- **Mean and Median:** The mean, median, and mode of a normal distribution are all equal and occur at the center of the distribution.

- **Standard Deviation:** The spread of the distribution is determined by the standard deviation. A smaller standard deviation results in a narrower distribution, while a larger standard deviation leads to a wider distribution.

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

Answer: Handling missing data is an important step in data preprocessing, as missing values can impact the quality and validity of our analysis and modeling. There are several approaches to handling missing data, and the choice depends on the nature of the data and the problem at hand. Here are some common techniques:

1. **Deletion:** We can simply remove rows or columns with missing values. However, this should be used cautiously, as it can lead to loss of valuable information.
2. **Imputation:** Imputation involves filling in missing values with estimated or predicted values. Some common imputation techniques include:
    - **Mean/Median Imputation:** Replace missing values with the mean or median of the available data for that variable. This is a simple method but can distort the distribution of the variable.
    - **Mode Imputation:** For categorical variables, you can replace missing values with the mode (most common value) of the variable.
    - **Regression Imputation:** Use regression models to predict missing values based on other variables. This can be more accurate if there's a strong relationship between the variable with missing values and other variables.
    - **K-Nearest Neighbors (KNN) Imputation:** Impute missing values using values from the k-nearest neighbors based on other features.
    - **Multiple Imputation:** Generate multiple plausible imputations to account for uncertainty. Each imputed dataset is analyzed separately, and results are combined.

12. What is A/B testing?

Answer: A/B testing, also known as split testing or bucket testing, is a controlled experiment used in marketing, product development, and other fields to compare two versions of a webpage, app feature, email, or any other element with the goal of determining which version performs better. The primary purpose of A/B testing is to make data-driven decisions by evaluating the impact of changes on user behavior and outcomes.

Here is how A/B testing works:

1. **Baseline Version (A):** The baseline version, often referred to as "A," is the current version that you want to compare against.
2. **Variant Version (B):** The variant version, often referred to as "B," is the modified version with one or more changes that you want to test.
3. **Random Assignment:** Users or participants are randomly assigned to either the baseline version (A) or the variant version (B) of the element being tested.
4. **Data Collection:** User interactions and outcomes, such as clicks, conversions, purchases, or any relevant metrics, are collected and recorded for both versions.
5. **Comparison and Analysis:** The collected data is then analyzed to determine whether the variant version (B) outperforms the baseline version (A) based on the chosen metrics. Statistical techniques are used to assess whether the observed differences are statistically significant.
6. **Decision-Making:** Based on the analysis, you can make informed decisions about which version is more effective in achieving the desired outcomes. If the variant version (B) performs better, you might consider implementing the changes in the production environment.

A/B testing allows businesses to test hypotheses, optimize user experiences, and make iterative improvements to their products or services. It's widely used in digital marketing to optimize landing pages, emails, ads, and more. The key to successful A/B testing is to ensure that the test is well-designed, the sample size is appropriate, and the results are statistically sound.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation of missing data is a common practice, but it has limitations such as distorting data distribution, reducing variability, impacting relationships, lacking accuracy, and overestimating certainty, making it less suitable in cases where the missing data mechanism is not completely random or when more advanced imputation methods could provide better results.

14. What is linear regression in statistics?

Answer: 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. The primary goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the observed data points and the predicted values generated by the linear equation.

In the case of simple linear regression, there's only one independent variable, and the linear equation takes the form:

$y=\beta 0+\beta 1x+\epsilon$

- y is the dependent variable.

- x is the independent variable.
- β0 is the intercept term.
- β1 is the coefficient of the independent variable.
- ϵ represents the error term or residual.

The coefficients β0 and β1 are estimated from the data to create the best-fitting line that minimizes the sum of squared residuals (the differences between actual and predicted values). This line represents the linear relationship between the variables.

In multiple linear regression, there are more than one independent variables, and the linear equation is extended to accommodate them:

$y=\beta0+\beta1x1+\beta2x2+\ldots+\beta pxp+\epsilon$

Linear regression is widely used for various purposes, including:

- **Prediction:** It can be used to predict the value of the dependent variable based on the values of the independent variables.
- **Inference:** It can help understand the relationship between variables and identify which variables have significant effects on the dependent variable.
- **Control:** In experimental settings, linear regression can be used to control for the effects of other variables and isolate the effect of the variable of interest.
- **Trend Analysis:** It's used in time-series analysis to identify trends and patterns in data.

Linear regression assumes that the relationship between variables is linear and that the errors are normally distributed and have constant variance. It's a foundational technique in statistics and serves as a building block for more complex regression and machine learning methods.

15. What are the various branches of statistics?

Answer: Statistics is a diverse field with several branches that cater to various aspects of data analysis, interpretation, and application. Some of the main branches of statistics include:

1. **Descriptive Statistics:** This branch involves summarizing and presenting data in a meaningful way, using measures like mean, median, mode, range, standard deviation, and graphical representations like histograms and bar charts.
2. **Inferential Statistics:** Inferential statistics deals with drawing conclusions and making predictions about populations based on samples. It includes hypothesis testing, confidence intervals, and regression analysis.
3. **Probability Theory:** Probability theory studies randomness and uncertainty. It provides a foundation for understanding and quantifying uncertainty in various statistical analyses.
4. **Biostatistics:** Biostatistics focuses on analyzing and interpreting data related to biological and medical sciences. It's used in clinical trials, epidemiology, and health research.
5. **Econometrics:** Econometrics applies statistical methods to economic data to test economic theories and analyze economic relationships.
6. **Social Statistics:** Social statistics deals with data related to social phenomena and human behavior. It's used in sociology, psychology, education, and other social sciences.
7. **Business Statistics:** Business statistics involves analyzing data related to business operations, market trends, and consumer behavior. It's used for decision-making and strategic planning.

8. **Environmental Statistics:** Environmental statistics focuses on analyzing environmental data to understand patterns, trends, and potential impacts on the environment.
9. **Statistical Computing:** This branch involves developing and using computational methods to analyze and interpret data. It includes software development, data visualization, and machine learning.
10. **Bayesian Statistics:** Bayesian statistics uses Bayes' theorem to update probabilities based on new evidence. It's used for modeling uncertainty and making predictions.
11. **Nonparametric Statistics:** Nonparametric methods do not assume a specific distribution for the data and are used when traditional parametric methods may not be appropriate.
12. **Multivariate Statistics:** Multivariate statistics deals with the analysis of data involving multiple variables. Techniques include principal component analysis, factor analysis, and cluster analysis.
13. **Spatial Statistics:** Spatial statistics analyzes data with a geographic component, considering spatial relationships and patterns.
14. **Time Series Analysis:** Time series analysis focuses on data that is collected over time, aiming to understand trends, seasonality, and patterns.

These are just a few of the many branches of statistics. Each branch caters to specific domains and applications, contributing to the broad field of statistics as a whole.