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

ans 1: 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

ans 2: 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

ans 3: C

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

ans 4: D

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

FLIP ROBO

ans 5: C

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

ans 6: 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

ans 7: 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

ans 8: 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

ans 9: C

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

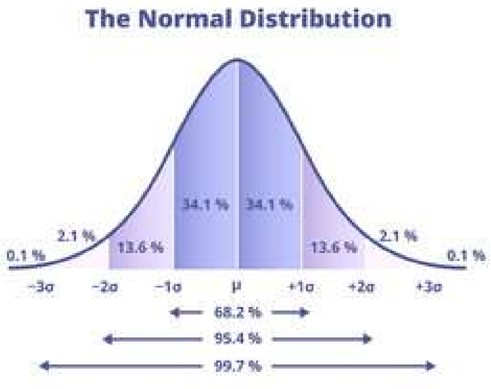
10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Q 10 to Q 15 in next pages

FLIP ROBO

Ans 10:

Normal distribution, also known as Gaussian distribution, is a probability distribution that is characterized by a bell-shaped curve.



It is a continuous probability distribution that is used to model a wide variety of natural phenomena, such as the distribution of heights or weights in a population, errors in measurement, and the sum of a large number of independent random variables.

The normal distribution is defined by two parameters: the mean (μ) and the standard deviation (σ). The mean is the center of the distribution, and the standard deviation measures the spread of the distribution. The shape of the normal distribution is determined by the mean and standard deviation, and it is symmetric around the mean.

The probability density function (PDF) of the normal distribution is given by the formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$f(x)$ = probability density function

σ = standard deviation

μ = mean

Ans 11:

Missing data can lead to biased or inaccurate results. The choice of imputation technique depends on the type and amount of missing data, the nature of the dataset, and the goals of the analysis .There are several ways to handle missing data in machine learning, including:

1. Complete Case Analysis (CCA): This approach involves removing all rows that contain missing values from the dataset. This method is simple, but it can lead to a loss of information in case the large number of rows is removed.

2. Mean/Mode/Median Imputation: This approach involves replacing missing values with the mean/mode/median value of the remaining data. This method is simple and easy to implement, but it can lead to biased estimates.

3. K-Nearest Neighbor (KNN) Imputation: This approach involves using the values of the K nearest neighbors of a missing value to estimate its value. This method is more accurate than mean imputation, but it can be computationally intensive and may not work well for high-dimensional datasets. KNN imputation may be more appropriate for small or moderate-sized datasets with missing values

4. Multiple Imputation: This approach involves creating multiple imputations of missing values and analyzing each imputed dataset separately. This method is more complex but can lead to more accurate results than single imputation methods. Multiple imputation is generally recommended for large datasets with many missing values.

5. Deep Learning-Based Imputation: This approach involves training a deep learning model to predict missing values based on the other features in the dataset. This method is more complex and computationally intensive, but it can lead to more accurate results than traditional imputation methods. Deep learning-based imputation is a promising approach for handling missing data, but it requires large amounts of data and computing resources.

Ans 12:

A/B testing is a statistical hypothesis testing method used to compare two versions of a product, website, or marketing campaign, to determine which one performs better. In A/B testing, two variants (A and B) are compared by randomly assigning users or subjects to either group A or group B, and measuring a chosen metric for each group. The goal is to determine whether the difference in the metric between the two groups is statistically significant or due to chance.

A/B testing typically involves the following steps:

1. Define the goal: The first step is to define the goal of the A/B test, such as increasing click-through rates, conversion rates, or revenue.
2. Create variants: Two or more variants are created, such as different versions of a website page or different email subject lines.
3. Randomly assign users: Users are randomly assigned to one of the variants, with an equal chance of being in either group.
4. Collect data: Data is collected on the chosen metric for each group, such as click-through rates or conversion rates.
5. Analyze results: The data is analyzed to determine if there is a statistically significant difference between the two groups, using statistical hypothesis testing methods.
6. Implement the winner: The variant that performs better is implemented, based on the results of the test.

A/B testing is widely used in marketing and product development to optimize website design, email campaigns, and product features. It is an effective way to test different ideas and determine what works best, based on data and statistical analysis.

Ans 13:

In most of the cases using Mean Imputation for missing data is not recommended. One of the main drawbacks of mean imputation is that it can introduce bias in the data, especially if the missing data is not missing at random. If the missing data is related to the outcome variable or other variables in the dataset, mean imputation can distort the relationships between variables and lead to incorrect conclusions.

Ans 14:

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. In simple linear regression, the relationship between the dependent variable Y and the independent variable X is modeled as a straight line:

$$Y = \beta 0 + \beta 1X + \varepsilon$$

where β0 is the intercept, β1 is the slope, X is the independent variable, ε is the error term, and Y is the dependent variable.

The goal of linear regression is to estimate the values of β0 and β1 that best fit the data, by minimizing the sum of squared errors between the predicted values and the actual values. The estimated coefficients can be used to make predictions of the dependent variable for new values of the independent variable.

Linear regression can be used for both prediction and inference, and is widely used in various fields, such as economics, finance, engineering, and social sciences.

Ans 15:

Statistics is a vast field that encompasses many sub-disciplines. Here is a brief overview of some of the main branches of statistics:

1. **Descriptive statistics**: Deals with the summary and presentation of data using measures such as mean, median, mode, variance, and standard deviation.
2. **Inferential statistics**: Deals with making inferences about a population based on a sample, using hypothesis testing, confidence intervals, and regression analysis.

3. **Probability theory**: Deals with the study of random phenomena and their properties, including probability distributions, expected values, and random variables.
4. **Econometrics**: Applies statistical methods to problems in economics, including forecasting, time-series analysis, and causal inference.
5. **Bayesian statistics**: Deals with the interpretation of probability as a degree of belief, and uses Bayes' theorem to update beliefs based on new evidence.
6. **Data mining**: Deals with the discovery of patterns and relationships in large datasets, using techniques such as clustering, classification, and association analysis.
7. **Machine learning**: Deals with the development of algorithms and models that can learn from data and make predictions or decisions, using techniques such as neural networks, decision trees, and support vector machines.
8. **Statistical computing**: Deals with the development of software and algorithms for statistical analysis, including programming languages such as R and Python, and software packages such as SAS and SPSS.
9. **Survey methodology**: Deals with the design, implementation, and analysis of surveys, including sampling techniques, questionnaire design, and data analysis.