WORKSHEET

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

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

3. Which of the following is incorrect with respect to use of Poisson distribution?
b) Modeling bounded count data

4. Point out the correct statement.
d) All of the mentioned

5. _____ random variables are used to model rates.
c) Poisson

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

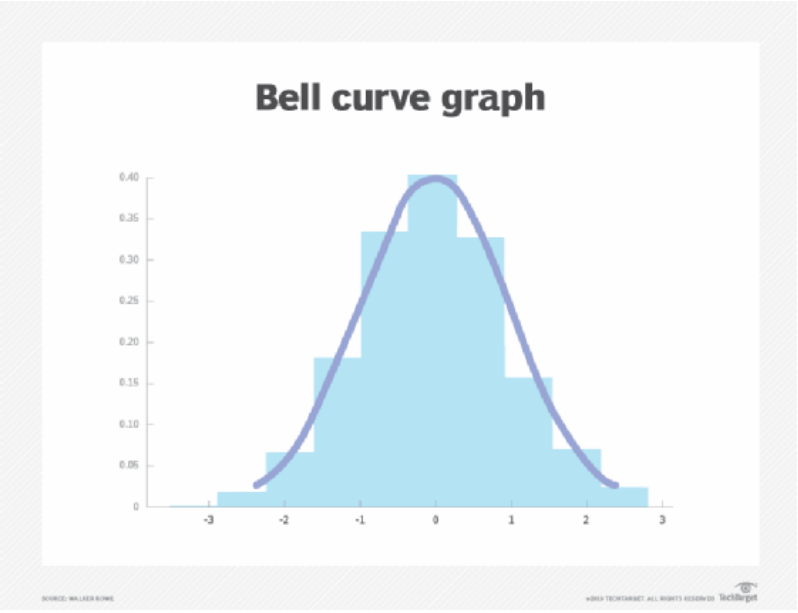
7. 1. Which of the following testing is concerned with making decisions using data?
b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0

9. Which of the following statement is incorrect with respect to outliers?
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?

A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, The middle of the range is also known as the mean of the distribution.



A normal distribution is a bell curve because of its flared shape. The precise shape can vary according to the distribution of the values within the population. The population is the entire set of data points that are part of the distribution.

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

There are varies method used in handle missing data, imputation techniques are used to commonly to fill the NAN values. The types of imputers are as follows,

1. Simple Imputer
2. K Nearest Neighbors
3. Iterative imputer

SimpleImputer:

SimpleImputer is a class in the sklearn.impute module that can be used to replace missing values in a dataset, using a variety of input strategies. Simple Imputer is designed to work with numerical data. SimpleImputer can be used as part of a scikit-learn Pipeline. The default strategy is “mean”, which replaces missing values with the median value of the column.

SimpleImputer class is used to impute / replace the numerical or categorical missing data related to one or more features with appropriate values such as following:

- Mean
- Median
- Mode
- Constant

K Nearest Neighbors:

KNN Imputer retains the most data compared to other techniques such as removing rows or columns with missing values. It replaces missing values with imputed values. KNN Imputer imputes missing values based on the nearest neighbors, which means it preserves the underlying relationships in the data. It takes into account the feature similarities between data points to estimate the missing values, making it more contextually relevant.

KNN Imputer is a non-parametric method, which means it does not make assumptions about the data’s distribution. It is suitable for both numeric and categorical data, making it versatile in handling various types of missing values.

The KNN Imputer allows you to specify the number of nearest neighbors to use for imputation

By using multiple neighboring data points to estimate missing values, KNN Imputer reduces the bias that may be introduced when using simple imputation techniques like mean or median imputation.

Iterative imputer:

Its implementation involves imputing missing values by modelling each feature as a function of other elements. The missing values are considered targets, and the remaining features are used to predict their values.

Each feature is imputed sequentially, one after the other, allowing prior imputed values to be used as part of a model in predicting subsequent features. It is iterative because this process is repeated multiple times, allowing ever improved estimates of missing values to be calculated as missing values across all features are estimated.

12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

A/B testing can help you evaluate the performance of your machine learning models in a real-world setting, compare different models or algorithms, and optimize your model parameters or features.

A/B testing machine learning models can pose some challenges that need to be addressed. These include dealing with noisy or incomplete data, accounting for confounding factors and interactions, and handling ethical and privacy issues. To address data quality, you may need to apply data cleaning, imputation techniques.

13. Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

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. It finds the best line that minimizes the differences between predicted and actual values.

The strength of any linear regression model can be assessed using various evaluation metrics. These evaluation metrics usually provide a measure of how well the observed outputs are being generated by the model.

The most used metrics are:

- Coefficient of Determination or R-squared (R2)
- Root Mean Squared Error (RSME) and Residual Standard Error (RSE)

1. **Coefficient of Determination or R-Squared (R2)**

   R-squared is a number that explains the amount of variation that is explained/captured by the developed model. It always ranges between 0 & 1. Overall, the higher the value of R-squared, the better the model fits the data.

2. **Root Mean Squared Error**

   The Root Mean Squared Error is the square root of the variance of the residuals. It specifies the absolute fit of the model to the data i.e. how close the observed data points are to the predicted values.

15. What are the various branches of statistics?

Basically, the statistical analysis is meant to collect and study the information available in large quantities. Statistics is a branch of mathematics, where computation is done over a bulk of data using charts, tables, graphs, etc.

Statistics have majorly categorized into two types:

- Descriptive statistics
- Inferential statistics

Descriptive statistics:

In this type of statistics, the data is summarized through the given observations. The summarization is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures.

Descriptive statistics are also categorized into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

Inferential Statistics:

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analyzed and summarized then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Importance of Statistics

- Statistics executes the work simply and gives a transparent picture of the work we do regularly.
- The statistical methods help us to examine different areas such as medicine, business, economic, social science and others.
- Statistics equips us with different kinds of organized data with the help of graphs, tables, diagrams and charts.
- Statistics helps to understand the variability of the data pattern in a quantitative way
- Statistics makes us understand the bulk of data in a simple way
- Statistics is the way to collecting accurate quantities data