STATISTICS WORKSHEET-3

1. Which of the following is the correct formula for total variation?
Total Variation = Residual Variation + Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called_____ outcomes.
Ans. binomial

3. How many outcomes are possible with Bernoulli trial?
Ans. 2

4. If Ho is true and we reject it is called
Ans. Type-I error

5. Level of significance is also called:
Ans. Power of the test

6. The chance of rejecting a true hypothesis decreases when sample size is:
Ans. Increase

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

8. What is the purpose of multiple testing in statistical inference?
Ans. All of the mentioned

9. Normalized data are centered at and have units equal to standard deviations of the original data
Ans. 0

10. What Is Bayes' Theorem?
In statistics and probability theory, the Bayes' theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

The theorem is named after English statistician, Thomas Bayes, who discovered the formula in 1763.
It is considered the foundation of the special statistical inference approach called the Bayes' inference.

Besides statistics, the Bayes' theorem is also used in various disciplines, with medicine and pharmacology as the most notable examples.
In addition, the theorem is commonly employed in different fields of finance. Some of the applications include but are not limited to,
modeling the risk of lending money to borrowers or forecasting the probability of the success of an investment.

11. What is z-score?
A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units.
The z-score is positive if the value lies above the mean, and negative if it lies below the mean.
It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution.

It allows researchers to calculate the probability of a score occurring within a standard normal distribution.
Enables us to compare two scores that are from different samples (which may have different means and standard deviations).

The formula for calculating a z-score is is $z = (x-\mu)/\sigma$, where x is the raw score, μ is the population mean, and σ is the population standard deviation.

12. What is t-test?
A t-test is a statistical test that is used to compare the means of two groups.
It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest,or whether two groups are different from one another.
A t-test can only be used when comparing the means of two groups.

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.
The t-test is one of many tests used for the purpose of hypothesis testing in statistics. Calculating a t-test requires three key data values. They include the difference between the mean values from each data set (called the mean difference), the standard deviation of each group, and the number of data values of each group.

13. What is percentile?
In statistics, percentiles are used to understand and interpret data.
The nth percentile of a set of data is the value at which n percent of the data is below it. In everyday life, percentiles are used to understand values such as test scores, health indicators, and other measurements.

For example, an 18-year-old male who is six and a half feet tall is in the 99th percentile for his height. This means that of all the 18-year-old males, 99 percent have a height that is equal to or less than six and a half feet.
An 18-year-old male who is only five and a half feet tall, on the other hand, is in the 16th percentile for his height, meaning only 16 percent of males his age are the same height or shorter.

Percentiles are used to understand and interpret data. They indicate the values below which a certain percentage of the data in a data set is found.

Percentiles can be calculated using the formula n = (P/100) x N, where P = percentile, N = number of values in a data set (sorted from smallest to largest), and n = ordinal rank of a given value.

Percentiles are frequently used to understand test scores and biometric measurements.

14. What is ANOVA?
Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts:

Systematic factors and Random factors.

The systematic factors have a statistical influence on the given data set, while the random factors do not.
Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

15. How can ANOVA help?
An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help to figure out if we need to reject the null hypothesis or accept the alternate hypothesis.
Basically, our testing groups to see if there's a difference between them.

SQL

1.Write SQL query to create table Customers.

```
create database ERD;
create table Customers(
   -> customerNumber int,
   -> customerName varchar(10),
   -> contactLastName varchar(10),
   -> contactFirstName varchar(10),
   -> phone int,
   -> addressLine1 varchar(10),
   -> addressLine2 varchar(10),
   -> city varchar(10),
   -> state varchar(10),
```

-> postalCode int,
-> country varchar(10),
-> saleRepEmployeeNumber int,
-> creditLimit int);

2.Write SQL query to create table Orders.

create table Orders
-> (orderNumber int,
-> orderDate varchar(10),
-> requiredDate varchar(10),
-> shippedDate varchar(10),
-> status varchar(10),
-> commnets varchar(10),
-> customerNumber int);

1. Write SQL query to show all the columns data from the Orders Table.

select * from orders;

2. Write SQL query to show all the comments from the Orders Table.

select commnets from orders;

3. Write a SQL query to show orderDate and Total number of orders placed on that date, from Orders table.

select orderDate from orders;

6.Write a SQL query to show employeNumber, lastName, firstName of all the employees from employees table.

select employeeNumber,lastName,firstName from employees;

7. Write a SQL query to show all orderNumber, customerName of the person who placed the respective order.

8. Write a SQL query to show name of all the customers in one column and salerepemployee name in another column.

select customerName ,saleRepEmployeeNumber from customers;.

9.Write a SQL query to show Date in one column and total payment amount of the payments made on that date from the payments table.

select paymentDate,amount from payments;.

10.Write a SQL query to show all the products productName, MSRP, productDescription from the products table.

select productName ,MSRP,productDescription from products;

11. Write a SQL query to print the employee number in one column and Full name of the employee in the second column for all the employees.

select employeenumber, concat(firstname,'',lastname) as fullname from employees;

12. Write a SQL query to print the orderNumber, customer Name and total amount paid by the customer for that order (quantityOrdered × priceEach).

MACHINE LEARNING

1. Which of the following is an application of clustering?
Ans. Biological network analysis

2. On which data type, we cannot perform cluster analysis?
Ans. Time series data

3. Netflix's movie recommendation system uses-
Ans. Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is-
Ans. The tree representing how close the data points are to each other

5. Which of the step is not required for K-means clustering?
Ans. None

6. Which is the following is wrong?
Ans. k-nearest neighbor is same as k-means

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?
Ans. 1, 2 and 3

8. Which of the following are true?
Ans. 1 only

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?
Ans. 2

10. For which of the following tasks might clustering be a suitable approach?
Ans.

11. Given, six points with the following attributes:
Ans. A

12. Given, six points with the following attributes:
Ans. B

13. What is the importance of clustering?

Ans. Clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome. Clustering is usually used to classify data into structures that are more easily understood and manipulated.

14. How can I improve my clustering performance?
Ans. K-means clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm.

When the data has overlapping clusters, k-means can improve the results of the initialization technique.

When the data has well separated clusters, the performance of k-means depends completely on the goodness of the initialization.

Initialization using simple furthest point heuristic (Maxmin) reduces the clustering error of k-means from 15% to 6%, on average.