STATISTICS WORKSHEET-3

1 B

2 C

3 A

4 A

5 B

6 B

7 B

8 D

9 A

10 Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

where:

$P(A) =$ The probability of A occurring

$P(B) =$ The probability of B occurring

$P(A|B) =$ The probability of A given B

$P(B|A) =$ The probability of B given A

$P\left(A \bigcap B\right)) =$ The probability of both A

11 Z-score indicates how much a given value differs from the standard deviation. The z-score measures exactly how many standard deviations above or below the mean a data point is. A positive z-score says the data point is above average. A negative z-score says the data point is below average.

$$Z = \frac{x - \mu}{\sigma}$$

Z = standard score

x = observed value

μ = mean of the sample

σ = standard deviation of the sample

12 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 If you want to compare more than two groups, or if you want to do multiple pairwise comparisons, use an ANOVA test.

The t test assumes

- are independent
- are (approximately) normally distributed
- have a similar amount of variance within each group being compared (a.k.a. homogeneity of variance)

if the data doesn't satisfy the following then nonparametric alternative to the t test, such as the Wilcoxon Signed-Rank test for data with unequal variances can be done.

When choosing a t test need to consider two things

- groups being compared come from a single population or two different populations
- whether you want to test the difference in a specific direction.

One-sample, two-sample, or paired t test are the different types of t test

- If the groups come from a single population (e.g., measuring before and after an experimental treatment), perform a **paired t test**. Eg: within-subjects design.
- If the groups come from two different populations (e.g., two different species, or people from two separate cities), perform a **two-sample t test** (a.k.a. independent t test). Eg: between-subjects design.
- If there is one group being compared against a standard value (e.g., comparing the acidity of a liquid to a neutral pH of 7), perform a **one-sample t test**
- If you only care whether the two populations are different from one another, perform a **two-tailed t test.**
- If you want to know whether one population mean is greater than or less than the other, perform a **one-tailed t test**.

13 A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. Eg: if a person scores 75 points on a test, and are ranked in the 85th percentile, it means that the score 75 is higher than 85% of the scores

14 The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples. If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1.

15 The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variable.

SQL

Q1 CREATE TABLE customers (

customerNumber int primary key NOT NULL,

customerName varchar(50) NOT NULL,

contactLastName varchar(15) NOT NULL,

contactFirstName varchar(30) NOT NULL,

phone int NOT NULL,

addressLine1 varchar(100) NOT NULL,

addressLine2 varchar(60) NOT NULL,

city varchar(25) NOT NULL,

state varchar(20) NOT NULL,

postalCode int NOT NULL,

country varchar(20) NOT NULL,

salesRepEmployeeNumber varchar(20) NOT NULL,

creditLimit int NOT NULL

);

Q2 CREATE TABLE orders (

orderNumber int NOT NULL PRIMARY KEY,

orderDate DATE NOT NULL,

requiredDate DATE NOT NULL,

shippedDate DATE NOT NULL,

status varchar(100) NOT NULL,

comments varchar(100),

customerNumber int NOT NULL,

FOREIGN KEY (orderNumber) REFERENCES orderdetails(orderNumber) ON

DELETE CASCADE

FOREIGN KEY (customerNumber) REFERENCES

customers(CustomerNumber)

ON DELETE CASCADE

);

DELETE CASCADE to specify whether you want rows deleted in a child table when corresponding rows are deleted in the parent table.

Q3 SELECT * from orders;

Q4 SELECT comments from orders;

Q5 SELECT orderDate, COUNT(orderNumber) from orders GROUP BY orderDate;

Q6 SELECT employeeNumber, lastName, firstName from employees;

Q7 SELECT orders.orderNumber, customers.customerName from orders. orders, customers. employees WHERE orders. customerNumber= customers. customerNumber

Q8 . SELECT customerName, CONCAT(firstName, ' ', lastName) from employees LEFT JOIN customers ON customers.salesRepEmployeeNumber=employees.employeeNumber;

Q9 SELECT paymentDate, sum(amount) from payments GROUP BY paymentDate;

Q10 SELECT productName, MSRP, productDescription from products;

Q11 SELECT products.productName, products.productDescription FROM products WHERE

MAX(COUNT(productCode));

Q12 SELECT customers.city, COUNT(orders.orderNumber) FROM customers LEFT

JOIN order ON customer.customerNumber = orders.customerNumber GROUP BY city

ORDER BY COUNT(orders.orderNumber) DESCS LIMIT 1;

Q13 SELECT state from customers where max(GROUP BY state);

Q14 SELECT employeeNumber, CONCAT(firstName,' ', lastName) AS FullName from employees;

Q15 SELECT orders.orderNumber,customers.customerName, (orderdetails .quantityOrdered * orderdetails. priceEach) AS TotalAmountPay FROM orders

LEFT JOIN customers ON orders.customerName = customer.customerNumber

LEFT JOIN orderdetails ON orderdetails.orderNumber = order.orderNumber ;

ML

1 d

2 d

3 c

4 b

5 d

6 c

7 d

8 a

9 a

10 b

11 a

12 a

13 Clustering is the process of arranging a group of objects in such a manner that the objects in the same group (which is referred to as a cluster) are more similar to each other than to the objects in any other group. Data professionals often use clustering in the Exploratory Data Analysis phase to discover new information and patterns in the data. As clustering is unsupervised machine learning, it doesn't require a labeled dataset. They play a wide role in applications like marketing economic research and weblogs to identify similarity measures, Image processing, and spatial research. They are used in outlier detections to detect credit card fraudulence.

14 performance can be improved by

- Merging neighboring clusters if the resultant clusters variance is less than threshold
- Isolating elements that are far if the cluster variance is above threshold
- moving some elements between neighboring clusters if it decreases the sum of squared errors