STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

1. What is central limit theorem and why is it important?

The central limit theorem states that the sampling distribution of the mean approaches a normal distribution, as the sample size increases. This fact holds especially true for sample sizes over 30.

Therefore, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean μ and standard deviation σ .

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution.

2. What is sampling? How many sampling methods do you know?

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.

There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified.

Random sampling is analogous to putting everyone's name into a hat and drawing out several names. Each element in the population has an equal chance of occuring. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained. Computer generated lists are often used with random sampling. You can generate random numbers using the TI82 calculator.

Systematic sampling is easier to do than random sampling. In systematic sampling, the list of elements is "counted off". That is, every kth element is taken. This is similar

to lining everyone up and numbering off "1,2,3,4; 1,2,3,4; etc". When done numbering, all people numbered 4 would be used.

Convenience sampling is very easy to do, but it's probably the worst technique to use. In convenience sampling, readily available data is used. That is, the first people the surveyor runs into.

Cluster sampling is accomplished by dividing the population into groups -- usually geographically. These groups are called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.

Stratified sampling also divides the population into groups called strata. However, this time it is by some characteristic, not geographically. For instance, the population might be separated into males and females. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

3. What is the difference between type1 and typeII error?

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.

There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified.

Random sampling is analogous to putting everyone's name into a hat and drawing out several names. Each element in the population has an equal chance of occuring. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained. Computer generated lists are often used with random sampling. You can generate random numbers using the TI82 calculator.

Systematic sampling is easier to do than random sampling. In systematic sampling, the list of elements is "counted off". That is, every kth element is taken. This is similar to lining everyone up and numbering off "1,2,3,4; 1,2,3,4; etc". When done numbering, all people numbered 4 would be used.

Convenience sampling is very easy to do, but it's probably the worst technique to use. In convenience sampling, readily available data is used. That is, the first people the surveyor runs into.

Cluster sampling is accomplished by dividing the population into groups -- usually geographically. These groups are called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.

Stratified sampling also divides the population into groups called strata. However, this time it is by some characteristic, not geographically. For instance, the population might be separated into males and females. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

The Type II error is denoted by β (beta) and is also termed as the beta error.

The null hypothesis is set to state that there is no relationship between two variables and the cause-effect relationship between two variables, if present, is caused by chance.

Type II error occurs when the null hypothesis is acceptable considering that the relationship between the variables is because of chance or luck, and even when there is a relationship between the variables.

As a result of this error, the researcher might end up believing that the hypothesis doesn't work even when it should.

4. What do you understand by the term Normal distribution?

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

5. What is correlation and covariance in statistics?

Correlation:-

Correlation is a statistical measure that indicates how strongly two variables are related.

It show whether and how strongly pairs of variables are related to each other.

Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.

In this variable are indirectly related to each other.

It gives the direction and strength of relationship between variables.

Covariance:-

It is the relationship between a pair of random variables where change in one variable causes change in another variable.

It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.

It is used for the linear relationship between variables.

It gives the direction of relationship between variables.

6. Differentiate between univariate ,Biavariate,and multivariate analysis.

Univarate Analysis:-

Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Since it's a single variable it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

You can think of the variable as a category that your data falls into. One example of a variable in univariate analysis might be "age". Another might be "height". Univariate analysis would not look at these two variables at the same time, nor would it look at the relationship between them.

Some ways you can describe patterns found in univariate data include looking at mean, mode, median, range, variance, maximum, minimum, quartiles, and standard deviation. Additionally, some ways you may display univariate data include frequency distribution tables, bar charts, histograms, frequency polygons, and pie charts.

Bivarate Analysis:-

Bivariate analysis is used to find out if there is a relationship between two different variables. Something as simple as creating a scatterplot by plotting one variable against another on a Cartesian plane (think X and Y axis) can sometimes give you a picture of what the data is trying to tell you. If the data seems to fit a line or curve then there is a relationship or correlation between the two variables. For example, one might choose to plot caloric intake versus weight.

Multivariate Analysis:-

Multivariate analysis is the analysis of three or more variables. There are many ways to perform multivariate analysis depending on our goals.

7. What do you understand by sensitivity and how would you calculate it?

A sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty. This technique is used within specific boundaries that depend on one or more input variables.

Sensitivity is calculated as :- A/(A+C) × 100 where, A = True positives C = False negatives

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You're basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment won't be repeatable and so has little use.

H0 is null hypothesis.

H1 is alternate hypothesis, the hypothesis we are interested in proving.

In a two-tailed test, the generic null(H0) and alternative hypotheses(H1) are the following:

Null(H0): The effect equals zero.

Alternative(H1): The effect does not equal zero.

9. What is quantitative data and qualitative data?

Quantitative data:-This data type is measured using numbers and values, making it a more suitable candidate for data analysis. Whereas qualitative is open for exploration, quantitative data is much more concise and close-ended. It can be used to ask the questions “how much” or “how many,” followed by conclusive information.

Qualitative data:-

Qualitative data is non-statistical and is typically unstructured or semi-structured. This data isn’t necessarily measured using hard numbers used to develop graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers.

Qualitative data can be used to ask the question “why.” It is investigative and is often open-ended until further research is conducted. Generating this data from qualitative research is used for theorizations, interpretations, developing hypotheses, and initial understandings.

10. How to calculate range and interquartile range?

Range:-

The Range is the difference between the lowest and highest values. Example: In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9. So the range is 9 − 3 = 6.

Interquartile rage:-

We can find the interquartile range or IQR in four simple steps:

Order the data from least to greatest

Find the median

Calculate the median of both the lower and upper half of the data

The IQR is the difference between the upper and lower medians

11. What do you understand by bell curve distribution ?

The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape.

A standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean.

12. Mention one method to find outliers.

Using Z-scores to Detect Outliers:-

Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean.

To calculate the Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation.

13. What is p-value in hypothesis testing?

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.

The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected.

A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

14. What is the Binomial Probability Formula?

The Binomial Probability distribution of exactly x successes from n number of trials is given by the below formula-

P (X) = nCx px qn – x Where,

n = Total number of trials

x = Total number of successful trials

p = probability of success in a single trial

q = probability of failure in a single trial = 1-p

15. Explain ANOVA and it's applications.

A common approach to figure out a reliable treatment method would be to analyse the days it took the patients to be cured. We can use a statistical technique which can compare these three treatment samples and depict how different these samples are from one another. Such a technique, which compares the samples on the basis of their means, is called ANOVA.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

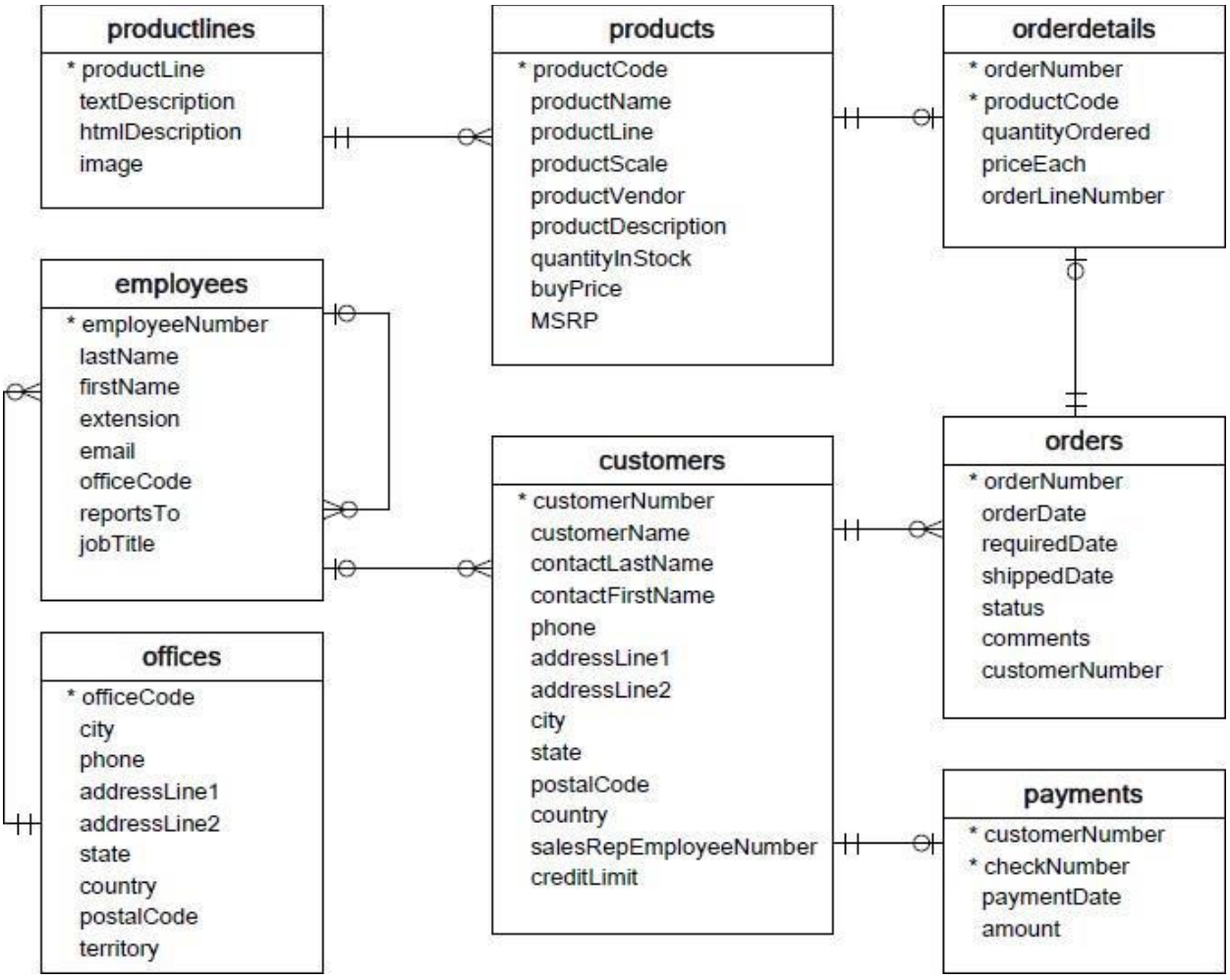
One real-life application of analysis of variance is the recommendation of a fertilizer against others for the improvement of a crop yield.

Anova can be in different fields of sciences, i.e. all the problems of testing more than three groups.

.

WORKSHEET 4 SQL

Refer the following ERD and answer all the questions in this worksheet. You have to write the queries using MySQL for the required Operation.



- **Customers**: stores customer's data.
- **Products**: stores a list of scale model cars.
- **Product Lines**: stores a list of product line categories.
- **Orders**: stores sales orders placed by customers.

- **Order Details**: stores sales order line items for each sales order.
- **Payments**: stores payments made by customers based on their accounts.
- **Employees**: stores all employee information as well as the organization structure such as who reports to whom.
- **Offices**: stores sales office data.

QUESTIONS:

1. Write a SQL query to show average number of orders shipped in a day (use Orders table).

```
select avg(orderNumber) from order where   status="shipped";
```

2. Write a SQL query to show average number of orders placed in a day.

```
select avg from order where orderDate="2022-5-12";
```

3. Write a SQL query to show the product name with minimum MSRP (use Products table).

```
select productName from products where MSRP=(select MIN(MSRP) from products);
```

ASSIGNMENT

4. Write a SQL query to show the product name with maximum value of stockQuantity.

select productName from products where quantityInStock=(select MAX (quantityInStock) from products);

5. Write a query to show the most ordered product Name (the product with maximum number of orders).

select p.productName from product as p INNER JOIN orderDetails as o on p.productCode=o.productCode where o.quantityOrdered=(select MAX.quantityOrdered);

6. Write a SQL query to show the highest paying customer Name.

select customerName from customers where creditLimit=(select max(creditLimit) from customers);

7. Write a SQL query to show cutomerNumber, customerName of all the customers who are from Melbourne city.

select customerNumber, customerName from customer where city="Melbourne";

8. Write a SQL query to show name of all the customers whose name start with “N”.

select customerName from customers where customerName like "N%";

9. Write a SQL query to show name of all the customers whose phone start with ‘7’ and are from city ‘Las Vegas’.

select customerName from customers where phone like "%7" and city="LasVegas";

10. Write a SQL query to show name of all the customers whose creditLimit < 1000 and city is either “Las Vegas” or ”Nantes” or “Stavern”.

select customerName from customers where creditLimit < 1000 and city="LasVegas" or "Nantes" or "Stavern";

11. Write a SQL query to show all the orderNumber in which quantity ordered <10.

select orderNumber from orderDetails where quantityOrdered < 10;

12. Write a SQL query to show all the orderNumber whose customer Name start with letter ‘N’**.**

select o.orderNumber from orders as o Inner Join cutomers as c ON c.customerNumber = o.customerNumber where c.customerName like "N%";

13. Write a SQL query to show all the customerName whose orders are “Disputed” in status.

select CustomerName from orders where status = "Disputed";

14. Write a SQL query to show the customerName who made payment through cheque with checkNumber starting with H and made payment on “2004-10-19”.

select c.customerName from customers as c Inner Join payment as p on c.customerNumber = p.customerNumber where p.chequeNumber like "H%" and p.paymentDate = "2004-10-19";

15. Write a SQL query to show all the checkNumber whose amount > 1000.

select checkNumber from payments where amount > 1000;

Machine Learning worksheet 4

1 In Q1 to Q7, only one option is correct, Choose the correct option:

1. The value of correlation coefficient will always be:
A) between 0 and 1 B) greater than -1
C) between -1 and 1 D) between 0 and -1
Ans-C

2. Which of the following cannot be used for dimensionality reduction?
A) Lasso Regularisation B) PCA
C) Recursive feature elimination D) Ridge Regularisation
Ans-C

3. Which of the following is not a kernel in Support Vector Machines?
A) linear B) Radial Basis Function
C) hyperplane D) polynomial
Ans-C

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
A) Logistic Regression B) Naïve Bayes Classifier
C) Decision Tree Classifier D) Support Vector Classifier
Ans-D

5. In a Linear Regression problem, ‘X’ is independent variable and ‘Y’ is dependent variable, where ‘X’ represents weight in pounds. If you convert the unit of ‘X’ to kilograms, then new coefficient of ‘X’ will be?
(1 kilogram = 2.205 pounds)
A) 2.205 × old coefficient of ‘X’ B) same as old coefficient of ‘X’
C) old coefficient of ‘X’ ÷ 2.205 D) Cannot be determined

Ans-C

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
A) remains same B) increases
C) decreases D) none of the above
Ans-B

7. Which of the following is not an advantage of using random forest instead of decision trees?
A) Random Forests reduce overfitting
B) Random Forests explains more variance in data then decision trees
C) Random Forests are easy to interpret
D) Random Forests provide a reliable feature importance estimate
Ans-B

In Q8 to Q10, more than one options are correct, Choose all the correct options:
8. Which of the following are correct about Principal Components?
A) Principal Components are calculated using supervised learning techniques
B) Principal Components are calculated using unsupervised learning techniques
C) Principal Components are linear combinations of Linear Variables.
D) All of the above
Ans-D

9. Which of the following are applications of clustering?

A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.
C) Identifying spam or ham emails
D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.
Ans-A,B,C,D

10. Which of the following is(are) hyper parameters of a decision tree?
A) max_depth B) max_features
C) n_estimators D) min_samples_leaf

Ans-A,B,D

Q10 to Q15 are subjective answer type questions, Answer them briefly.

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

Q1 represents the 25th percentile of the data. Q2 represents the 50th percentile of the data. Q3 represents the 75th percentile of the data. If a dataset has 2n / 2n+1 data points, then Q1 = median of the dataset. Q2 = median of n smallest data points. Q3 = median of n highest data points.

IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

12. What is the primary difference between bagging and boosting algorithms?

Bagging is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.

Boosting is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

13. What is adjusted R2 in linear regression. How is it calculated?

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

Adjusted R Squared = 1 – [((1 – R2) * (n – 1)) / (n – k – 1)]

Where:

n – Number of points in your data set. k – Number of independent variables in the model, excluding the constant

14. What is the difference between standardisation and normalisation?

In normalisation Minimum and maximum value of features are used for scaling, In standarisation Mean and standard deviation is used for scaling.

Normalisation is used when features are of different scales. Standarization is used when we want to ensure zero mean and unit standard deviation.

In normalisation Scales values between [0, 1] or [-1, 1]. standarisation is not bounded to a certain range.

Normalisation is really affected by outliers. Standarisation is much less affected by outliers.

Scikit-Learn provides a transformer called MinMaxScaler for Normalization. Scikit-Learn provides a transformer called StandardScaler for standardization.

Normalisation transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Standarisation translates the data to the mean vector of original data to the origin and squishes or expands.

Normalisation is useful when we don't know about the distribution. Standarisation is useful when the feature distribution is Normal or Gaussian.

Normalisation is a often called as Scaling Normalization. Standarisation is a often called as Z-Score Normalization.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

Advantage:-

An advantage of using this method is that we make use of all data points and hence it is low bias.

Disadvatage:-

The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point. If the data point is an outlier it can lead to higher variation.