STATISTICS ASSIGNMENT 4

Q1. What is central limit theorem and why is it important?

Ans:

The central limit theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size gets larger no matter what the shape of the population distribution.

The central limit theorem is important because it is used in hypothesis testing to calculate intervals.

Q2. What is sampling? How many sampling methods do you know?

Ans:

The sampling method or sampling technique is the process of studying the population by gathering information and analyzing the data. It is based on the data where the sample space is enormous.

The several different sampling techniques available, and they can be subdivided into two groups.

All these methods of sampling may involve specifically targeting had or approaching to reach group.

Types of Sampling Method:

There is different sampling technique available to get relevant results from the population. The two different types of sampling methods are:

- Probability Sampling:
  It involves random selection, allowing us to make strong statistical inference about the whole group.

- Non-Probability Sampling:
  It involves non-random selection based on convenience or other criteria, allowing us to easily collect data.

Q3. What is the difference between type1 and type II error?

Ans:

Type 1 error:

- A type 1 error is also known as a false positive and occurs when a researcher incorrectly rejects a true null Hypothesis.
- A type 1 error also known as False Positive.
- The Probability that we will make a type 1 error is designated 'α' (alpha) Therefore type 1 error is also known as alpha error.
- Type 1 error is associated with rejecting the null hypothesis.
- It is caused by luck or chance
- The probability of Type1 error reduces with lower values of ('α') since the lower value makes it difficult to reject null hypothesis.
- Type 1 error are generally considered more serious.
- It can be reduced by decreasing the level of significance.
- The probability of type1 error is equal to the level of significance.
- It happens when the acceptance levels are set too lenient

Type II Error:

- A type II error does not reject the null hypothesis, even though the alternative hypothesis is the true state of nature. In other words, a false finding is accepted as true.
- A type II error also known as False Negative. It is also known as false null hypothesis.
- Probability that we will make a type II error is designated 'B'(Beta). Therefore, type II error Is also known as Beta Error.
- Type II error equals to the statistical power of a test.
- Type II error equals to the statistical power of a test, the probability 1-B is called the statistical power of the study

- Type II error is associated with rejecting the alternative hypothesis.

- It is caused by a smaller sample size or a less powerful test.
- The probability of Type ii error reduces with higher values of ('α') since the higher vlue makes it easier to reject the null hypothesis.
- Type II error gives less preference.

- It can be reduced by increasing the level of significance.
- The probability of type ii error is equal to one minus the power of the test.
- It happens when the acceptance levels are set too stringent.

Q4. What do you understand by the term Normal distribution?

Ans:

The Normal Distribution Is the most common type of Distribution assumed in Technical stock market analysis and in other type of Statistical Analysis. The Standard normal Distribution has two parameters: The mean and the standard Deviation.

- Normal distribution also known as the Gaussian Distribution or Bell Curve.
- The normal distribution where mean, median and mode is equal to zero, The standard Deviation is 1, Skewness is Zero.
- Normal distribution is symmetrical, but not all symmetrical distribution are normal.
- The normal distribution model is important in statistics and is key to the central limit theorem.

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

- *x* = value of the variable or data being examined and f(x) the probability function
- μ = the mean
- σ = the standard deviation.

Q5. What is correlation and covariance in statistics?

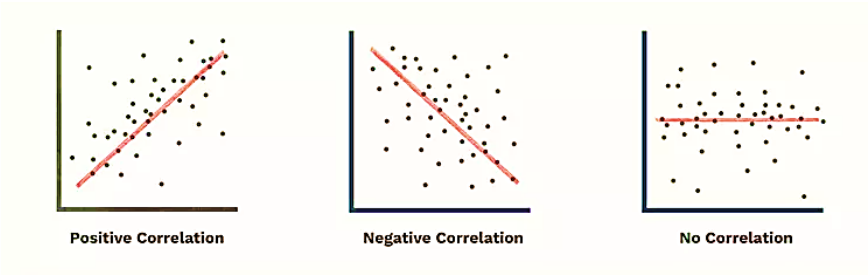
Ans:

Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It's a common tool for describing simple relations without making a statement about cause and effect. It is used to test relationships between quantitative variables or categorical variables. In other words, It's a measure of how

things are related. The study of how variables are correlated is called correlation analysis.

Points:

- A correlation coefficient is a way to put a value to the relationship.
- Correlation Coefficients have a value of between -1 and 1.
- A"0" means there is no relationship between the variables.
- While -1 or 1 means that there is perfect negative or positive correlation.



Correlation is of three type:

- Simple Correlation:
  A single Number expresses the deh=gree to which two variables are related.

- Partial Correlation
  When one variables effects are removed the correlation between two variables is revealed in partial correlation.

- Multiple Correlation
  A statistical technique that uses two or more variables to predict the values of one variable.

Correlation Formula

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

Where,

n: Quantity of Information

Σx: Total of the First Variable Value

Σy: Total of the Second Variable Value

Σxy: Sum of the Product of & Second Value

Σx2: Sum of the Squares of the First Value

Σy2: Sum of the Squares of the Second Value

Covariance:

It's a measure of the relationship between two random variables and to what extent, They change together, Or It defines the changes between the two Variable, such that change in one variable is equal to change in other variable. This is the property of a function of maintaining its form when the variables are linearly transformed.Covariance is measured in units, which are calculated by multiplying the units of the two variables.

Types of Covariance.

Covariance can have both postive and negative values. Based on this, it has two types:

- Poitive covariance
- Negative Covariance

1. Posiive Covariance:

If the covariance for any variable is positive that means both the variables move in the same direction. Hence they have similar behaviour. That means if the variable values(greater or lesser) then they are said to be in positive covariance.

2. Negative Covariance:

If the covariance for any two variable is negative that means both variable moves in the opposite direction. It is the oposite case of positive covariance, where greater values of one variable correspond to less values of another variable and vice versa.

Covariance Formula:

Population Covariance Formula

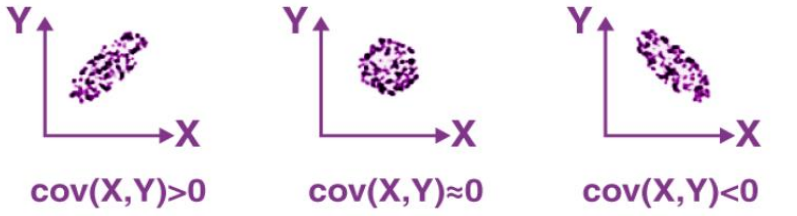
$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - y)}{N - 1}$$

Where,

- Yi= Data values of y
- Xi= Data values of x
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of Data Values.



Q6. Differentiate between univariate, Bivariate, and multivariate analysis

Ans:

1. Univariate:

It summarizes only one variable at a time. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

The example of a univariate data can be height.

2. Bivariate:

It compares two variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.

Example of Bivariate: temperature and ice cream sales in summer season.

3. Multivariate:

It compares more than two variables. It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).

Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

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

Ans:

- Sensitivity Analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions.
- Model is also referred to as a what if or simulation analysis.
- Sensitivity analysis can be used to help make predictions in the share prices of publicly traded companies or how interest rates affect bond prices.
- Sensitivity analysis allows for forecasting using historical, true data.
- While sensitivity analysis determines how variables impact single event.
- Sensitivity analysis is used in the business world and in the field of economics.

Calculate Sensitivity Analysis:

Sensitivity analysis is often performed in analysis software, and Excel has built in functions to help perform the analysis. In general, sensitivity analysis is calculated

by leveraging formulas that reference different input cells. For example, a company may perform NPV analysis using a discount rate of 6%. Sensitivity analysis can be performed by analyzing scenarios of 5%, 8%, and 10% discount rates as well by simply maintaining the formula but referencing the different variable values.

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

Ans:

Hypothesis or Significance Testing is a mathematical model for testing a claim, idea or hypothesis about a parameter of interest in a given population set. Calculations are performed on selected samples to gather more decisive information about the characteristics of the entire population.

Steps: Define the Hypothesis:

Usually, the report value (or the claim statistics) is stated as the hypothesis and assumed True.

Ex: A students in the school score an average of 7 out of 10 in exam.

This stated description constitutes the "Null Hypothesis (H0)"and is assumed to be True – the way a defendant in a jury trail is presumed innocent until proven guilty by the evidence presented in court. Similarly, hypothesis testing starts by stating and assuming a null hypothesis. And then the process determines whether the assumption is likely to be true or false.

The important point to note is that we are testing the null hypothesis because there is an element of doubt about its validity. Whatever information that is against the stated null hypothesis is captured in the Alternative Hypothesis(H1)

Ex: Student score an average that is not equal to 7

Then the alternative hypothesis is True.

Two Tailed Hypothesis Testing:

In two tails, the test samples is checked to be greater or less than a range of values in a Two-Tailed test, implying that the critical distribution area is two sided.

Example:

Suppose,

Ho: mean=50 and

H1: mean ≠ 50 (The mean can be greater than or less than 50 but not equal to)

Example:

The average height of students in a batch is 100 cm and the standard deviation is 15. However, one teacher believes that this has changed, so he/she decides to test the height of 75 random students in the batch. The average height of the sample comes out to be 105.

The steps for performing hypothesis testing are:

• Specify the Null(H0) and Alternate(H1) hypothesis

• Choose the level of Significance(α)

• Find Critical Values

• Find the test statistic

• Draw conclusion

➢ Specify the Null(H0) and Alternate(H1) hypothesis for Two Tailed Hypothesis Testing:

• Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. In this example, it will be the average height of students in the batch is 100

$$H0: \mu = 100$$

• Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed. In this example, teacher believes (Claims) that the actual value has

changed. He/she doesn't know whether the average has gone up or down, but he/she believes that it has changed and is not 100 anymore.

$$H1: \mu \neq 100$$

If the alternate hypothesis is written with a ≠ sign that means that we are going to perform a two-tailed test because chances, are it could be more than 100 or less than 100 which makes it two-tailed as an alternate hypothesis is always written with a ≠ or < or > sign.

Q9. **What is quantitative data and qualitative data?**

Ans:

Quantitative data:

This Data is Based on the Quantity ex: How Many, how much, How Often

- Quantitative data is anything that can be counted or measured; it refers to numerical data.

Example:

1. How much revenue did the company make in 2019?
2. How often does a certain customer group use online banking?

Qualitative Data:

Those which can't be Measured or Counted based on numerical values. Ex: Why? How?

Qualitative data also refers to the words or labels used to describe certain characteristics or traits.

For example, describing the sky as blue or labelling a particular ice cream flavor as vanilla.

Q10. How to calculate range and interquartile range?

Ans:

Range: To calculate range we need to find the max and minimum value of the variable and calculated as

Range=(Xmax-Xmin)

Ex: [2,5,8,10,3] 10-2=8 is range

Interquartile Range:

In descriptive statistics, the interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data.

Q1= 25% percent of points are found when arranged in ascending order.

Q2= 50% percentage or median is considered the second Quartile.

Q3=75% of data points are found when arranged in increasing order.

Formula

IQR=Upper Quartile-Lower Quartile

Q11. What do you understand by bell curve distribution?

Ans:

A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side. Bell curves are visual representations of normal distribution, also called Gaussian distribution.

A normal distribution curve, when graphed out, typically follows a bell-shaped curve, hence the name. While the precise shape can vary according to the distribution of the population, the peak is always in the middle and the curve is always symmetrical.

Bell curves are useful for quickly visualizing a data set's mean, mode and median because when the distribution is normal, the mean, median and mode are all the same.

The long tail refers to the part of the bell curve that stretches out in either direction. If the diagram above represents a population under study, the fat area under the bell curve is where most of the population falls.

Q12. Mention one method to find outliers.?

Ans:

Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.

There are four ways to identify/detect outliers:

- Sorting method.
- Data visualization method.
- Statistical tests (z scores)
- Interquartile range method

Z-scores:

Z-scores can quantify the unusualness of an observation when our 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.

Formula of Z-score:

Zscore= Data Point- Mean Population/Standard Deviation

Q 13. What is p-value in hypothesis testing?

Ans: A p-value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis.

The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage. For example, a p value of 0.0254 is 2.54%. This means there is 2.54% chance your results could be random. That's pretty tiny. On the other hand, a large p value of 0.9(90%) means your results have 90 probability of being completely random and not due to anything in your experiment. Therefore, the smaller the p-value, the more important our results.

Q14. What is the Binomial Probability Formula?

Ans: The binomial distribution forms the base for the famous binomial test of statistical importance. A test that has a single outcome such as success/Failure is also called a Bernoulli trail or Bernoulli experiment, and a series of outcomes is called Bernoulli process. Consider an experiment where each time a question is asked for a yes/No with a series of n experiments. Then in the binomial probability distribution, the Boolean-valued outcome the success/yes/true/one is represented with probability p and the failure/no/false/zero with probability q(q=1-p). In a single experiment when n=1, the binomial distribution is called a Bernoulli Distribution.

Formula for binomial distribution:

$$P(X) = {}_nC_x p^x (1-p)^{n-x}$$

n= the number of experiments

X=0,1,2,3,4….

p=Probability of success in single experiment

q= Probability of failure in a single experiment(1-p)

Q15. Explain ANOVA and its applications?

Ans:

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 variable 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.

Types:

1. **One Way ANOVA**
   It is also known as one factor ANOVA. Here, we are using one criterion variable (or called as a factor) and analyze the difference between more than two sample groups.

   Suppose in glass industry, we want to compare the variation of three batches(glass) for their average weight (factor).

2. **Two Way ANOVA**
   Here we are using two independent variable (factors) and analyze the difference between more than two sample groups. Similarly, we want to compare the variation of three batches of glass. w.r.t: weight and hardness (two factors)

THE FORMULA FOR ANOVA IS: F=MST/MSE

Where:

F= ANOVA coefficient

MST= MEAN SUM OF SQUARES DUE TO TREATMENT

MSE= MEAN SUM OF SQUARES DUE TO ERROR