STATISTICS WORKSHEET-4

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

Central limit theorem states that, given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population.

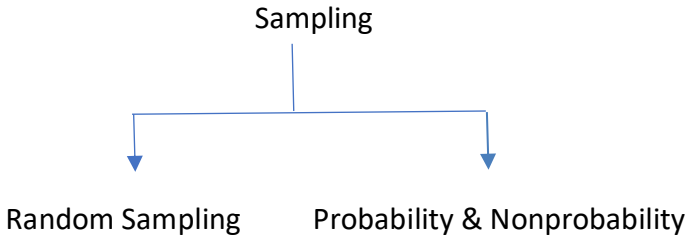
In other words, adding up all the means from all the samples, and finding the average and that average will be the actual population mean.

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, as we will see in the next section.

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

Sampling is a statistical procedure that is concerned with the selection of the individual observation; it helps us to make statistical inferences about the population.

Types of sampling:



Random sampling:

In data collection, every individual observation has equal probability to be selected into a sample. In random sampling, there should be no pattern when drawing a sample.

Probability and non-probability sampling:

Probability sampling is the sampling technique in which every individual unit of the population has greater than zero probability of getting selected into a sample. Non-probability sampling is the sampling technique in which some elements of the population have no probability of getting selected into a sample.

3. What is the difference between type I and type II error?

Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true. Type I error is equivalent to false positive.

Type II error is the error that occurs when the null hypothesis is accepted when it is not true. Type II error is equivalent to a false negative.

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.

Properties of Normal Distribution are as follows.

1. Unimodal -one mode
2. Symmetrical -left and right halves are mirror images
3. Bell-shaped -maximum height (mode) at the mean
4. Mean, Mode, and Median are all located in the centre
5. Asymptotic

5. What is correlation and covariance in statistics?

Correlation: Correlation is considered or described as the best technique for measuring and for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.

Covariance: In covariance two items vary together and it‘s a measure that indicates the extent to which two random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

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

Univariate analyses are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and the analysis can be referred to as univariate analysis.

The **Bivariate** analysis attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.

Multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

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

Sensitivity is used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.).

Sensitivity is Predicted True events/ Total events.
True events here are the events which were true and model also predicted them as true.

Calculation:

Seasonality = (**True Positives**) / (**Positives in Actual Dependent Variable**)

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

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter.
Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a
larger population, or from a data-generating process.

What is H0 and H1?
H1: The hypothesis that we are interested in proving. Null hypothesis: H0: The complement of the alternative hypothesis.
This is the probability of falsely rejecting the null hypothesis.

What is H0 and H1 for two-tail test?
Our null hypothesis is that the mean is equal to x. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x

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

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.

Quantitative Data – Quantitative data is statistical and is typically structured in nature – meaning it is more rigid and defined. This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

10. How to calculate range and interquartile range?

The formula to calculate the **range** is:

$$R = H - L$$

- R = range
- H = highest value
- L = lowest value

Interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

11. What do you understand by bell curve distribution?

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the centre is a mirror image of the left side.

The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis).

For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve.

12. Mention one method to find outliers.

by using z score and IQR method we can treat outliers and finding we can easily use box plot to identify the outliers

13. What is p-value in hypothesis testing?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called the Null Hypothesis.
Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value (≥ 0.05) indicates strength for the null

hypothesis which means we can accept the null Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,
High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

14. What is the Binomial Probability Formula?

P_{x} = {n \choose x} p^{x} q^{n-x}

P = binomial probability
x = number of times for a specific outcome within n trials
{n \choose x} = number of combinations
p = probability of success on a single trial
q = probability of failure on a single trial
n = number of trials

15. Explain ANOVA and it's applications.

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.

Applications of ANOVA –

- The ANOVA test is the initial step in analysing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.

- 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. The distribution of all possible values of the F statistic is the F-distribution. This is a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.