STATISTICS WORKSHEET-4

1) **What is central limit theorem and why is it important?**

Answer - The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

The central limit theorem is useful when analyzing large data sets because it allows one to assume that the sampling distribution of the mean will be normally-distributed in most cases. This allows for easier statistical analysis and inference. For example, investors can use central limit theorem to aggregate individual security performance data and generate distribution of sample means that represent a larger population distribution for security returns over a period of time.

2) **What is sampling? How many sampling methods do you know?**

Answer – Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate the 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.

Types of Sampling Methods –

1) Probability Sampling – It is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.

2) Non-Probability Sampling – Here, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

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

Answer- Type I and Type II errors are subjected to the result of the null hypothesis. In case of type I or type-1 error, the null hypothesis is rejected though it is true whereas type II or type-2 error, the null hypothesis is not rejected even when the alternative hypothesis is true. Both the error type-i and type-ii are also known as “false negative”. A lot of statistical theory rotates around the reduction of one or both of these errors, still, the total elimination of both is explained as a statistical impossibility.

Type I Error

A type I error appears when the null hypothesis (H0) of an experiment is true, but still, it is rejected. It is stating something which is not present or a false hit. A type I error is often called a false positive (an event that shows that a given condition is present when it is absent). The type I error significance level or rate level is the probability of refusing the null hypothesis given that it is true. It is represented by Greek letter α (alpha) and is also known as alpha level. Usually, the significance level or the probability of type I error is set to 0.05 (5%), assuming that it is satisfactory to have a 5% probability of inaccurately rejecting the null hypothesis.

Type II Error

A type II error appears when the null hypothesis is false but mistakenly fails to be refused. It is losing to state what is present and a miss. A type II error is also known as false negative (where a real hit was rejected by the test and is observed as a miss), in an experiment checking for a condition with a final outcome of true or false.

A type II error is assigned when a true alternative hypothesis is not acknowledged.

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

Answer - 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 graphical form, the normal distribution appears as a "bell curve".

The normal distribution model is important in statistics and is key to the Central Limit Theorem (CLT). This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance).

5) What is correlation and covariance in statistics?

Answer - Covariance is a statistical tool that is used to determine the relationship between the movements of two random variables. When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.

$$\text{Covariance} = \frac{\sum (\text{Return}_{ABC} - \text{Average}_{ABC}) * (\text{Return}_{XYZ} - \text{Average}_{XYZ})}{(\text{Sample Size}) - 1}$$

Covariance Formula.

Where:

- xi = a given x value in the data set
- xm = the mean, or average, of the x values
- yi = the y value in the data set that corresponds with xi
- ym = the mean, or average, of the y values

6) Differentiate between univariate, Bivariate and multivariate analysis.

Answer –

Univariate data – This type of data consists of **only one variable**. 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.

Bivariate data – This type of data involves **two different 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 data can be temperature and ice cream sales in summer season.

Multivariate data – When the data involves **three or more variables**, it is categorized under multivariate. 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.

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

Answer - 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 analysis is used in the business world and in the field of economics. It is commonly used by financial analysts and economists and is also known as a what-if 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.

8) **What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test**?
Answer - A statistical hypothesis is an assertion or conjecture concerning one or more populations. To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population. Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis.

Hypothesis testing is formulated in terms of two hypotheses:
• H0: the null hypothesis;
• H1: the alternate hypothesis.

The hypothesis we want to test is if H1 is "likely" true. So, there are two possible outcomes:

• Reject H0 and accept H1 because of sufficient evidence in the sample in favor or H1;

• Do not reject H0 because of insufficient evidence to support H1.

2 tailed hypothesis - A two-tailed hypothesis test is designed to show whether the sample mean is significantly greater than and significantly less than the mean of a population. The two-tailed test gets its name from testing the area under both tails (sides) of a normal distribution

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

Answer - Quantitative data refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it's quantitative data. Quantitative data can tell you "how many," "how much," or "how often"—for example, how many people attended last week's webinar? How much revenue did the company make in 2019? How often does a certain customer group use online banking?

To analyze and make sense of quantitative data, you'll conduct statistical analyses.

Unlike quantitative data, qualitative data cannot be measured or counted. It's descriptive, expressed in terms of language rather than numerical values.

Researchers will often turn to qualitative data to answer "Why?" or "How?" questions. For example, if your quantitative data tells you that a certain website visitor abandoned their shopping cart three times in one week, you'd probably want to investigate why—and this might involve collecting some form of qualitative data from the user. Perhaps you want to know how a user feels about a particular product; again, qualitative data can provide such insights. In this case, you're not just looking at numbers; you're asking the user to tell you, using language, why they did something or how they feel.

Qualitative data also refers to the words or labels used to describe certain characteristics or traits—for example, describing the sky as blue or labeling a particular ice cream flavor as vanilla

10) **How to calculate range and interquartile range?**
Answer - The range is the difference between the largest and smallest values in a data set.

The interquartile range (IQR) is the difference between the upper and lower quartiles.

For example: {2, 3, 5, 5, 7, 8, 9, 11, 12, 12, 50}

The range is 50 - 2 = 48

To find IQR, first find the median (the middle value) which is 8.

Then find the median of the numbers less then 8 and the numbers greater than 8. These give the lower quartile (5) and the upper quartile (12). The IQR is 12 - 5 = 7.

The IQR is resistant to outliers, such as the 50 in this data set. While that value makes the range large, it makes no difference to the IQR if that number is 50 or 12.

11) What do you understand by bell curve distribution?

Answer - A bell curve is a common type of distribution for a variable, also known as the normal distribution. The term "bell curve" originates from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.

The highest point on the curve, or the top of the bell, represents the most probable event in a series of data (its mean, mode, and median in this case), while all other possible occurrences are symmetrically distributed around the mean, creating a downward-sloping curve on each side of the peak. The width of the bell curve is described by its standard deviation.

12) Mention one method to find outliers

Answer - Outlier Detection using Z Score

The difference between any value in a set of data and the mean of that data, when measured in standard deviation units is simply called the **Z-score**, also known as the **standard score**.

Z-score may be negative value, positive value or 0. Z-score is 0 means, that data point is identical to mean of that data set. Positive value indicating the corresponding value above the mean. And negative value indicating the corresponding value below the mean.

We can simply calculate Z-score using following formula,

$$z = (x - \mu) / \sigma$$

- **x** is the raw score
- **μ** is the mean of the population
- **σ** is the standard deviation of the population

The Z-score is one method that can be used to detect outliers. Here it is assumed that the data set exists in a **Gaussian distribution**, also known as the **Normal distribution**. So, this method is most suitable for a data set with a normal distribution.

Usually z-score =3 is considered as a cut-off value to set the limit. Therefore, any z-score greater than +3 or less than -3 is considered as outlier which is pretty much similar to standard deviation method.

13) What is p-value in hypothesis testing?

Answer - The p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

P-values are used in hypothesis testing to help decide whether to reject the null hypothesis

14) What is the Binomial Probability Formula?

Answer - In the binomial probability, the number of successes X in 'n' trials of a binomial experiment is called a binomial random variable. The probability distribution of the random variable X is called a binomial distribution, and is given by the formula as below:

P(X)=Cnxpxqn−xP(X)=Cxnpxqn−x

Where n is the number of trials, x is 0, 1, 2..., n, p is the probability of success in a single trial, q is the probability of failure in a single trial and the value of q is 1-p. P(X) gives the probability of successes in n binomial trials.

The combination formula is Cnx=n!x!(n−x)!Cxn=n!x!(n−x)!.

15) Explain ANOVA and it's applications

Answer - ANOVA stands for Analysis of Variance. One-Way Analysis of Variance tells you if there are any statistical differences between the means of three or more independent groups.

You might use Analysis of Variance (ANOVA) as a marketer, when you want to test a particular hypothesis. You would use ANOVA to help you understand how your different groups respond, with a null hypothesis for the test that the means of the different groups are equal. If there is a statistically significant result, then it means that the two populations are unequal (or different).

The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variables (such as the first example: age, sex, income). When you understand how each independent variable's mean is different from the others, you can begin to understand which of them has a connection to your dependent variable (landing page clicks), and begin to learn what is driving that behaviour.