STATISTICS WORKSHEET-4

1. What is central limit theorem and why is it important?
   Ans. The Central Limit Theorem states that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not normally distributed. 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 tells us that no matter what the distribution of the population is, the shape of the sampling distribution will approach normality as the sample size (N) increases which is usually greater than 30.
   This is useful, as the research never knows which mean in the sampling distribution is the same as the population mean, but by selecting many random samples from a population the sample means will cluster together, allowing the research to make a very good estimate of the population mean. Thus, as the sample size (N) increases the sampling error will decrease.

2. What is sampling? How many sampling methods do you know?
   Ans. 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.
   So, to get accurate results or the results that can estimate the population well, the sampling technique should be chosen wisely. There are two types of sampling methods that can be used-
   i. 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. There are four types of probability sampling techniques:
      a. Simple random sampling
      b. Systematic sampling
      c. Stratified sampling
      d. Cluster sampling
   ii. Non-probability sampling: In non-probability sampling, 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.
      a. Convenience sampling
      b. Quota sampling
      c. Purposive sampling
      d. Snowball sampling

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

S.No.	BASIS FOR COMPARISON	TYPE I ERROR	TYPE II ERROR
1	Meaning	Type I error refers to non-acceptance of hypothesis which ought to be accepted.	Type II error is the acceptance of hypothesis which ought to be rejected.
2	Equivalent to	False positive	False negative
3	What is it?	It is incorrect rejection of true null hypothesis.	It is incorrect acceptance of false null hypothesis.
4	Represents	A false hit	A miss

5	Probability of committing error	Equals the level of significance.	Equals the power of test.
6	Indicated by	Greek letter 'α'	Greek letter 'β'

4. What do you understand by the term Normal distribution?
   Ans. 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".
   In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3. For all normal distributions, 68.2% of the observations will appear within plus or minus one standard deviation of the mean; 95.4% of the observations will fall within +/- two standard deviations; and 99.7% within +/- three standard deviations. This fact is referred to as the "empirical rule,".

5. What is correlation and covariance in statistics?
   Ans. Covariance measures how the two variables move concerning each other and is an extension of the concept of variance. It can take any value from -∞ to +∞. A positive number signifies positive covariance and denotes a direct connection. Effectively this means that an increase in one variable would also lead to a corresponding increase in the other variable, provided other conditions remain constant. On the other hand, a negative number signifies negative covariance, which denotes an inverse relationship between the two variables. Though covariance is perfect for defining the type of relationship, it is not good for interpreting its magnitude.
   Correlation is a step ahead of covariance as it quantifies the relationship between two random variables. In simple terms, it is a unit measure of how these variables change concerning each other (normalized covariance value).
   Positive correlation: Two variables are considered to have a positive correlation if they are directly proportional. That is, if the value of one variable increases, then the value of the other variable will also increase. A perfect positive correlation holds a value of “1”.
   Negative correlation: A perfect negative correlation holds a value of “-1” which means that, as the value of one variable increases, the value of the second variable decreases (and vice versa).

6. Differentiate between univariate ,Biavariate,and multivariate analysis.
   Ans. Univariate Analysis: It 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. Some ways you can describe patterns found in univariate data include looking at mean, mode, median, range, variance, maximum, minimum, quartiles, and standard deviation. Some ways that we display univariate data include frequency distribution tables, bar charts, histograms, frequency polygons, and pie charts.
   Bivariate Analysis: It refers to the exploratory data analysis between two variables. The variables can be either numeric or categorical. Bivariate analysis helps in studying the relationship between two variables, and if the two variables are related, we can comment on the strength of association. Some methods to do Bivariate analysis include Scatter Plat, Regression Plot, Chi-Squared Test & ANOVA.
   Multivariate Analysis: It refers to the statistical procedure for analyzing the data involving more than two variables. This can be used to analyze the relationship between dependent and independent variables. Multivariate analysis has various applications in clustering, feature selection, Multiple Regression Analysis, root-cause analysis, hypothesis testing, dimensionality reduction, etc.

7. What do you understand by sensitivity and how would you calculate it?
   Ans. Sensitivity is used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions. It is a way to predict the outcome of a decision given a certain range of variables. By creating a given set of variables, an analyst can determine how changes in one variable

affect the outcome. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
   Ans. Hypothesis testing is a tool for making statistical inferences about the population data. It is an analysis tool that tests assumptions and determines how likely something is within a given standard of accuracy. Hypothesis testing provides a way to verify whether the results of an experiment are valid. Hypothesis testing uses sample data from the population to draw useful conclusions regarding the population probability distribution. It tests an assumption made about the data using different types of hypothesis testing methodologies. The hypothesis testing results in either rejecting or not rejecting the null hypothesis. Hypothesis testing is formulated in terms of two hypothesis:
   H0: the null hypothesis - There's no effect in the population. A null hypothesis is a statistical hypothesis in which there is no significant difference exist between the set of variables. It is the original or default statement, with no effect.
   H1: the alternate hypothesis - There's an effect in the population. A statistical hypothesis used in hypothesis testing, which states that there is a significant difference between the set of variables. It is used to show that the observations of an experiment are due to some real effect. It indicates that there is a statistical significance between two possible outcomes.
   While the null hypothesis is the hypothesis, which is to be actually tested, whereas alternative hypothesis gives an alternative to the null hypothesis.

   A two-tailed test, in statistics, is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values. It is used in null-hypothesis(H0) testing and testing for statistical significance. If the sample being tested falls into either of the critical areas, the alternative hypothesis(H1) is accepted instead of the null hypothesis.
   If the mean of the sample is 18.
   H0: Null Hypothesis: mean = 18
   H1: Alternative Hypothesis: mean <> 18 (This is what we want to prove.)

9. What is quantitative data and qualitative data?
   Ans. Quantitative Data - 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".
   Qualitative Data - Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code. Qualitative data are data about categorical variables (e.g. what type). Qualitative data can be used to ask the question "why."

10. How to calculate range and interquartile range?
    Ans. Calculating Range - To calculate the range, you need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum). The range only takes into account these two values and ignore the data points between the two extremities of the distribution.
    Calculating Interquartile Range - To calculate these two measures, you need to know the values of the lower and upper quartiles. The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The interquartile range is the difference between upper and lower quartiles. i.e. IQR = Q3 – Q1

11. What do you understand by bell curve distribution ?
    Ans. 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.
    Ans. 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.
    The further away an observation's Z-score is from zero, the more unusual it is. A standard cut-off value for finding outliers are Z-scores of +/-3 or further from zero. As a rule of thumb, values with a z score greater than 3 or less than –3 are often determined to be outliers.

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. Graphically, the p value is the area in the tail of a probability distribution. It's calculated when you run hypothesis test and is the area to the right of the test statistic (if you're running a two-tailed test, it's the area to the left and to the right).
    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. The smaller the p-value, the more likely you are to reject the null hypothesis.
    A p-value less than 0.05 is typically considered to be statistically significant, in which case the null hypothesis should be rejected. A p-value greater than 0.05 means that deviation from the null hypothesis is not statistically significant, and the null hypothesis is not rejected.
    - A small p (≤ 0.05), reject the null hypothesis. This is strong evidence that the null hypothesis is invalid.
    - A large p (> 0.05) means the alternate hypothesis is weak, so you do not reject the null.

14. What is the Binomial Probability Formula?
    Ans. Binomial Distribution Formula:
    $P(x) = nCx \cdot px\ (1 - p)n-x$
    $P(x:n,p) = nCxx\ px\ (1-p)n-x$ Or $P(x:n,p) = nCx\ px\ (q)n-x$
    $P(x) = nCx \cdot px\ (1 - p)n-x$
    - n = Total number of events
    - r (or) x = Total number of successful events.
    - p = Probability of success on a single trial.
    - $nCr = [n!/r!(n-r)]!$
    - 1 – p = Probability of failure.

15. Explain ANOVA and it's applications.
    Ans. An ANOVA test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance. Another Key part of ANOVA is that it splits the independent variable into 2 or more groups. For example, one or more groups might be expected to influence the dependent variable while the other group is used as a control group and is not expected to influence the dependent variable.
    Types of ANOVA Tests:-
    i. A one-way ANOVA (analysis of variance) has one categorical independent variable (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable. The one-way ANOVA test for differences in the means of the dependent variable

ii. A two-way ANOVA (analysis of variance) has two or more categorical independent variables (also known as a factor), and a normally distributed continuous (i.e., interval or ratio level) dependent variable. A two-way ANOVA is also called a factorial ANOVA.

The formula for ANOVA is F = variance caused by treatment/variance due to random chance.

The t-test determines whether two populations are statistically different from each other, whereas ANOVA tests are used when an individual wants to test more than two levels within an independent variable.

i. A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis,
ii. A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis.

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.