STATISTICS WORKSHEET-4

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

Ans: The Central Limit Theorem 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 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.

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

Ans: Sampling means selecting the group that you will actually collect data from in your research. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students. In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

There are two types of sampling methods: Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group. Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Q3. What is the difference between type1 and typeII error?

Ans: Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false. When the null hypothesis is true but mistakenly rejected, it is type I error. As against this, when the null hypothesis is false but erroneously accepted, it is type II error.

Q4. 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".

Q5. What is correlation and covariance in statistics?

Ans: Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related.

Both covariance and correlation measure the relationship and the dependency between two variables. Covariance indicates the direction of the linear relationship between variables while correlation measures both the strength and direction of the linear relationship between two variables.

Q6. Differentiate between univariate ,Biavariate,and multivariate analysis.

Ans: Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

Univariate 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.

Bivariate 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.

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

Q7. What do you understand by sensitivity and how would you calculate it?

Ans: Sensitivity analysis is an analysis technique that works on the basis of what-if analysis like how independent factors can affect the dependent factor and is used to predict the outcome when analysis is performed under certain conditions.

The formula for sensitivity analysis is basically a financial model in excel where the analyst is required to identify the key variables for the output formula and then assess the output based on different combinations of the independent variables. Mathematically, the dependent output formula is represented as, $Z = X2 + Y2$

Q8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Ans: Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. The test provides evidence concerning the plausibility of the hypothesis, given the data. Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

In hypothesis testing there are two mutually exclusive hypotheses; the Null Hypothesis (H0) and the Alternative Hypothesis (H1). One of these is the claim to be tested and based on the sampling results (which infers a similar measurement in the population), the claim will either be supported or not.

A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x.

Q9. What is quantitative data and qualitative data?

Ans: Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

Q10. How to calculate range and interquartile range

Ans: The IQR describes the middle 50% of values when ordered from lowest to highest. To find the 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.

The interquartile range formula is the first quartile subtracted from the third quartile: IQR = Q3 – Q1.

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.

Q12. Mention one method to find outliers.

Ans: There are four ways to identify outliers:

1. Sorting method.
2. Data visualization method.
3. Statistical tests (z scores)
4. Interquartile range method.

Q13. What is p-value in hypothesis testing?

Ans: 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.

Q14. What is the Binomial Probability Formula?

Ans: Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment). If the probability of success on an individual trial is p, then the binomial probability is $nCx \cdot px \cdot (1-p)n-x$.

Q15. Explain ANOVA and it's applications

Ans: Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

ANOVA checks the impact of one or more factors by comparing the means of different samples. We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not. Another measure to compare the samples is called a t-test. When we have only two samples, t-test and ANOVA give the same results.