STATISTICS WORKSHEET – 4

1. What is central limit theorem and why is it important?
   Ans: 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 important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases.

2. What is sampling? How many sampling methods do you know?
   Ans: Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger 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.

3. 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.

   Type I error tends to assert something that is not really present, i.e. it is a false hit. On the contrary, type II error fails in identifying something, that is present, i.e. it is a miss.

   Greek letter ‘α’ indicates type I error. Unlike, type II error which is denoted by Greek letter ‘β’.

4. What do you understand by the term Normal distribution?
   Ans: A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

5. What is correlation and covariance in statistics?
   Ans: Correlation is a statistical measure that expresses the extent to which two variables are linearly related i.e.,meaning they change together at a constant rate. It is a common tool for describing simple relationships without making a statement about cause and effect.

   Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable.

6. 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**.

7. What do you understand by sensitivity and how would you calculate it?
   Ans: The sensitivity of a test is its ability to determine the patient cases correctly. To estimate it, we should calculate the proportion of true positive in patient cases.

   Mathematically, this can be stated as:

   Sensitivity = TP TP + FN.

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
   Ans: Hyphothesis testing is formulated in terms of two hypothesis:
   H0 : the null hypothesis.
   H1 : the alternate hypothesis.

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

- Reject H0 and accept H1 because of sufficient evidence in the sample in favour or H1.
- Do not reject H0 because of insufficient evidence of support H1.

9. 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.

10. How to calculate range and interquartile range?
    Ans: **Range** is the difference between the highest and lowest values in the data set.

    Interquartile range (IQR): Is the difference between the upper and lower quartiles.

11. What do you understand by bell curve distribution ?
    Ans: A bell curve refers to the graphical representation of normal probability distribution. The underlying standard deviations of this distribution from the median or the highest point of the curve give it the shape of a curved bell.

12. Mention one method to find outliers.
    Ans: Z-score Method

13. 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. The smaller the p value, the more likely you are to reject the null hypothesis.

14. What is the Binomial Probability Formula?

Ans: $\mathbf{P(X) = {}^{n}C_{x}p^{x}q^{n-x}}$

P = binomial probability

X = number of times for a specific outcome within n trials

${}^{n}C_{x}$ = 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.

Ans: ANOVA means Analysis of variance. ANOVA is a statistical method that separates observed variance data into different components to use for additional tests. 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 t-test. When we have only two samples, t-test and ANOVA give the same results.