STATISTICS WORKSHEET - 4

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

Ans. The central limit theorem states that if we have a large population's data and we are taking n numbers of random samples from the population then the distribution of the sample mean of all the n samples, will be approximately normally distributed as the sample size gets larger. But we have to satisfy one condition that is, The sample size should be equal to or greater than 30.

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

Ans. When you conduct the research for finding some insights from the group of people. It's rarely possible to collect the data for every person in that group. Instead, you select some samples. Here we can say that sample is the group of people who will actually participate in the research. To draw a valid conclusion, we need to decide first, in which way we will select the sample, which can represent the group as a whole. There are two types of sampling methods:

1) **Probability sampling** – It means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce a result that is representative of the whole population, Probability sampling is the most valid choice.

2) **Non-probability sampling** – In this technique, individuals are selected based on non-random criteria, and not every individual has a chance to be selected.
This type of sampling is easier and cheaper to access, but it has a higher risk of sampling bias. This sampling does not aim to test a hypothesis about the broad population but to develop an initial understanding of a small or under-researched population.

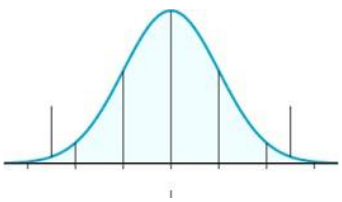
3. What is the difference between type1 and type II errors?

Ans. A Type I error (Alpha)(false-positive) means rejecting the null hypothesis when it's actually true. A Type II error (Beta)(false-negative) means not rejecting the null hypothesis when it's actually false.

Example-
Type I error (false positive): the test result says you have coronavirus, but you actually don't.
Type II error (false negative): the test result says you don't have coronavirus, but you actually do.

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

Ans. A normal distribution is also called Gaussian distribution. In this distribution, the mean, median, and mode are all equal to one another.
In the normal distribution, the mean is zero, the standard deviation is 1 and it has zero skewness. It is visually depicted as the "bell curve".



5. What are correlation and covariance in statistics?

Ans. Covariance – It is a measure of the relationship between the variability of two variables i.e. It measures the degree of changes in the variables when one variable changes will there be a similar change in the other variable?

"OR"

It is the relation between the pair of random variables where a change in one variable causes a change in another variable.

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

- xi = data value of x
- yi= data value of y
- x̄ = mean of x
- ȳ = mean of y ● N = number of data values.

Correlation – Correlation gives a better understanding of covariance. Correlation tells us how the variables are correlated to each other. It is also called the Pearson correlation coefficient.

$$\rho_{xy} = \frac{Con\,(r_x, r_y)}{\sigma_x \sigma_y}$$

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

Ans. Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Bivariate analysis is used to find out if there is a relationship between two different variables. Multivariate analysis is the analysis of three or more variables.

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

Ans. Sensitivity is the percentage of true positives (e.g. 90% sensitivity = 90% of people who have the target disease will test positive and the rest 10% are false negative). Specificity is the percentage of true negatives (e.g. 90% specificity = 90% of people who do not have the target disease will test negative and rest 10% are false positive).

Sensitivity=[TP/(TP+FN)]×100

Specificity=[TN/(FP+TN)]×100

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

Ans. Hypothesis testing is a statistical method that is used in making a statistical decision using experimental data.

"OR"

Hypothesis testing is basically an assumption that we make about the population parameter. It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
H0 - Null Hypothesis (Null hypothesis is a general given statement)

H1 – Alternative Hypothesis (Alternative hypothesis is used in hypothesizing that is contrary to the null hypothesis).

The two-sample t-test (Two tail test) compares the mean of two independent groups in order to determine the mean of two different variables are identical or not.
Here,

H0 – Two variables are independent.
H1 – Two variables are dependent.

9. What are quantitative data and qualitative data?

Ans. Quantitative data can be counted, measured, and expressed using numbers. It gives the hard facts. Whereas qualitative data is descriptive and have categories and it is used to gain an understanding of human behavior, intentions, attitudes, experience, etc., based on the observation and the interpretation of the people.

10. How to calculate range and interquartile range?

Ans. To calculate the range, you need to find the largest observed value of a variable (maximum) and subtract the smallest observed value (minimum).to calculate the range we need maximum value and minimum value.

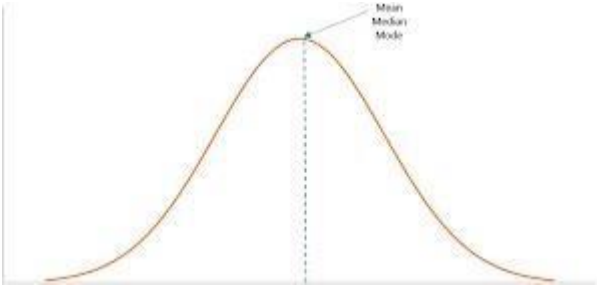
Range(r) = Max – Min

To calculate the interquartile range we need the 25th percentile value and the 75th percentile value.

IQR = Q3 – Q1
Q1 = 25th percentile value
Q3 = 75th percentile value

11. What do you understand by bell curve distribution?

Ans. A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell. The top of the curve shows the mean, mode, and median of the data collected. Its standard deviation depicts the bell curve's relative width around the mean.



12. Mention one method to find outliers.

Ans. Z -Score method – Z score is the method to find out the outliers present in the data set. For example, a z-score of 2.5 indicates that the data point is 2.5 standard deviations away from the mean. Usually z-score =3 is considered a cut-off value to set the limit. Therefore, any z-score greater than +3 or less than -3 is considered an outlier.

13. What is the p-value in hypothesis testing?

Ans. The p-value or calculated probability is the probability of finding the observed/extreme results when the null hypothesis of a given study is true.
If,

- P-value > 0.05 – Null hypothesis is correct and accepted and alternative hypothesis is rejected.
- P-value < 0.05 – Null hypothesis is rejected and the alternative hypothesis is accepted.

14. What is the Binomial Probability Formula?

Ans.

$$P_x = \binom{n}{x} p^x q^{n-x}$$

15. Explain ANOVA and its applications.

Ans. The Anova test allows a comparison of two or more than two groups at the same time it helps to determine whether a relationship exists between them or not. Anova is used to compare the difference of means among more than 2 groups. It does this by looking at variation in the data and where that variation is found. Anova compare the amount of variation between groups with the amount of variation within the group.