STATISTICS WORKSHEET-4

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

If we take different large samples from a population then the mean of all the samples become identical to its population mean.

It is foremost important as this theorem allows us to use normal distributions for large samples from other non-normal distributions.

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

Sampling is a statistical methodology that uses a portion of a total population to represent the full population, i.e. a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.

I know four types of sampling methods and they are as following-

- **Simple Random Sampling**
- **Systematic Sampling**
- **Stratified Sampling**
- **Cluster Sampling**

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

Type-I (false-positive) error occurs when we reject the actual true null hypothesis and **Type-II** (false-negative) error occurs when we accept the false null hypothesis.

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

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. The resulting curve is symmetrical about the mean and forms a bell-shaped distribution.

5. What is correlation and covariance in statistics?

Correlation and covariance are two statistical concepts used to determine the relationship between two random variables. Correlation defines how a change in one variable will impact the other, while covariance defines how two items vary together, i.e. how the data points differ from the mean.

The range for correlation is (-1, 1) where A positive correlation means that the linear relationship is positive, and the two variables increase or decrease in the same direction. A negative correlation is just the opposite, wherein the relationship line has a negative slope and the variables change in opposite directions i.e., one variable decreases while the other increases. Zero correlation simply means that the variables behave very differently and thus, have no linear relationship.

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

Univariate
This type of data consists of only one variable.

Bivariate
This type of data involves two different variables.

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

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

Let's understand the term sensitivity with an example-

Sensitivity is like- Out of all the real positive examples, how many are predicted as positive.

Basically it indicates that- what is the percent of the true positive predicted with respect to total actual positive samples.

Sensitivity= TP/ (TP+FN)

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

Hypothesis is the predetermined formal procedures used in statistics to determine whether condition should be accepted or rejected.

H0- Null Hypothesis (Statement In favor)

H1- Alternate Hypothesis (Statement in Against)

In two tailed test will compare the mean of the two variables and if the mean are identical and p-value will be greater than 0.05 then will accept the null hypothesis otherwise reject it.

H0- $p>0.05$(identical mean) – will accept the null hypothesis

H1- $p<0.05$(unequal mean) - will accept the Alternate hypothesis

9. What is quantitative data and qualitative data?

Quantitative is basically numeric representation of data like Discrete (Whole Number) and Continuous data e.g. Weight, Age, and Height etcetera. Whereas qualitative is categorical or nominal representation of the data e.g. Gender, Colors, Week day etcetera.

10. How to calculate range and interquartile range?

Range is basically the difference b/w the highest value and the lowest value of a dataset.

The formula for finding the interquartile range is basically the difference between third quartile and second quartile i.e. **(Q3-Q1)**. Equivalently, the interquartile range is the region between the 75th and 25th percentile i.e. (75 – 25) = 50% of the data.

11. What do you understand by bell curve distribution?

The term bell curve is used to describe the mathematical concept called normal distribution, sometimes referred to as Gaussian distribution. "Bell curve" refers to the bell shape that is created when a line is plotted using the data points for an item that meets the criteria of normal distribution.

In a bell curve, the center contains the greatest number of a value and, therefore, it is the highest point on the arc of the line.

12. Mention one method to find outliers.

The **Box plots** might be foremost important while finding-out the outliers. This type of chart highlights minimum and maximum values, the range, the median, and the interquartile range for our dataset so that we can easily find out the outliers by seeing it.

13. What is p-value in hypothesis testing?

A p value is used in hypothesis testing to help us 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 one should reject the null hypothesis.

A small p (≤ 0.05), reject the null hypothesis.

A large p (> 0.05) means the alternate hypothesis is weak, so we do not reject the null hypothesis.

14. What is the Binomial Probability Formula?

$P(x) = nCx * px (1 − p) n-x$

Where,

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.

We use Analysis of Variance (ANOVA) test when we have two or more than two numeric variables in the dataset and then we use to find out the variance between the variables column and among the variables. It is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.

ANOVA are getting heavily used in **Healthcare** and **Pharmaceutical** industry.