FLIP ROBO

STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

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

   Ans – It states that in case of large sample size the data will always be normally distributed regardless of the data type. According to the central limit theorem, the mean of a sample of data will be closer to the mean of the overall population. It is important to assume while making any hypothesis that the distribution is normal and because of this theorem we can assume that the sample distribution is normal regardless of the population distribution which helps us to make findings. Because of CLT we can take advantage of some statistical techniques where the distribution is required to be normal distribution.

2. What is sampling? How many sampling methods do you know?
   Ans – Sampling is a process or technique of selecting elements from a large group of data on a condition. It helps to understand the whole data and make findings out of it. There are 4 methods of sampling.
   1. Random Sampling – when all the elements of the group have equal probability of being chosen for a sample and samples are being chosen randomly is called random sampling.
   2. Systematic Sampling – systematic sampling follows certain conditions of being got selected. Systematic sampling includes certain strategy and randomness while choosing samples.
   3. Stratified Sampling – in this type the population is divided into groups based on some features and then samples would be pickup from each group on random basis.
   4. Clustering Sampling – clustering sampling is similar to stratified sampling. In this type of sampling technique the samples are being groped in by some features or characteristics where every samples has equal probability of selection.

3. What is the difference between type1 and typeII error?
   Ans – The Type I error refers to False Positive result where we reject the Null hypothesis where it is actually true and Type II error refers to False Negative result where we fail to reject a null hypothesis that is actually false.

4. What do you understand by the term Normal distribution?
   Ans – Normal distribution is a continuous probability distribution wherein values lie in a symmetrical fashion mostly situated around the mean and it creates a bell like shape.

5. What is correlation and covariance in statistics?
   Ans – covariance refers to the systematic relationship between two random variables. It could be positive or negative where change in one variable will lead to change in another. Covariance can be measure from – to + infinite, bigger the covariance the strong the relationship will be. Correlation is a measure that determines the degree to which two or more random variables move in sequence.

6. Differentiate between univariate, Biavariate, and multivariate analysis.
   Ans – Univariate - this type of data consists of only one variable.
   Biavariate – 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?
   Ans – Sensitivity is the metric that evaluates a model's ability to predict true positives of each available category.

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
   Ans – Hypothesis testing is a type of statistical analysis in which we put our assumptions about a population parameter to the test. To conclude the hypothesis, we have two assumption the one is the actual test and the other is the opposite of the test. H0 is called as Null Hypothesis which states that the event will not occur and H1 is called as Alternative Hypothesis is the opposite of Null hypothesis.

9. What is quantitative data and qualitative data?
   Ans – Quantitative data is data that can be counted or measured in numerical values. The two main types of quantitative data are discrete data and continuous data. Qualitative data is defined as the data that approximates and characterizes. Qualitative data can be observed and recorded. This data type is non-numerical in nature.

10. How to calculate range and interquartile range?
    Ans – To calculate the range we need to identified the highest value in the data and subtract that from the lowest value of the data. Interquartile range also refers to IQR which is a difference between quartile 1 and quartile 3.

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.

12. Mention one method to find outliers.
    Ans – Interquartile range method is one of the efficient methods to find outliers. To find outliers we need to first calculate 1^{st} and 3^{rd} quartile and IQR (the difference between these two quartiles) and with these we can find outliers which are not allowed above and below the specific level.

13. What is p-value in hypothesis testing?
    Ans – p-value is a statistical measurement used to validate a hypothesis against observed data. It measures the probability of obtaining the observed results, assuming that the null hypothesis is true. If the p-value is lower then there is a greater statistical significance of the observed value.

14. What is the Binomial Probability Formula?
    Ans –
    P(x) = (n/x)p^x q^n-x

15. Explain ANOVA and it's applications.
    Ans – ANOVA means Analysis of variance which is a collection of statistical models and their associated estimation procedures used to analyze the differences among means. ANOVA can be used to test hypothesis and it can be used to understand various data point and find observation.