FLIP ROBO

WORKSHEET-4(STATISTICS)

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

Ans: Central limit theorem states that, given a sufficiently large sample size the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population.

Central limit theorem is important 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 the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population.

There are two sampling methods, probability sampling and non-probability sampling. 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 allowing you to easily collect data.

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

Ans: **Type I error** is the error caused by rejecting a null hypothesis when it's true.

Type II error is the error that occurs when null hypothesis is accepted when it's not true.

Type I is equivalent to false positive and **Type II** is equivalent to false negative

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

Ans: **Normal distribution** also called 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. Normal distribution will appear as bell curve.

5. **What is correlation and covariance in statistics?**

Ans: Correlation is a statistical measure that indicates how strongly two variables are related or it measures the strength and direction of linear relationship

Covariance is a measure of how much two random variables vary together or it is the linear relationship between two variables..

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

Ans: Univariate analysis provides summary statistics for each field in the raw dataset or summary only on one variable.

Bivariate analysis is performed to find relationship between each variable in the dataset and the target variable of interest.

Multivariate analysis is performed to understand interactions between different fields in the dataset or finding interactions between more than two variables.

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

Ans: Sensitivity or true positive rate measures the proportion of positives that are correctly identified i.e proportion of those who have some condition who are correctly identified as having the condition. Sensitivity = A / (A+C) X100

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

Ans: Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. H1 is the alternative hypothesis-case that we are interested in proving. H0 is the null hypothesis. It is the complement of alternative hypothesis. Hypothesis testing is formulated in terms of two hypotheses H0 and H1.

9. **What is quantitative data and qualitative data?**

Ans: Quantitative data are data about numeric variables.

Qualitative data are measures of types and maybe represented by name, symbol etc. Quantitative methods allow us to test a hypothesis by systematically collecting and analysing data

Qualitative method allows you to explore ideas and experience in depth.

10. How to calculate range and interquartile range?
Ans: Range is the spread of data from lowest to highest value in the distribution. Range is the difference between highest and the lowest values. Interquartile range is the first quartile subtracted from third quartile. IQR=Q3-Q1

11. What do you understand by bell curve distribution?
Ans: Bell curve is used to describe a normal probability distribution whose underlying standard deviation from the mean create curved bell shape. A standard deviation is a measurement used to quantify the variability of data dispersion in a set of given values around the mean.

12. Mention one method to find outliers.
Ans: The most effective way to find outliers is by using interquartile range. IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR

13. What is p-value in hypothesis testing?
Ans: P Value or calculated probability is probability of finding the observed or more extreme, results when null hypothesis of a study question is true. The definition of extreme depends on how hypothesis is tested.

14. 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.
P(x) = •C. p (1 — p)n-r

15. Explain ANOVA and its applications.
Ans: Analysis of variance (ANOVA) is a statistical technique used to check if the means of two or more groups are significantly different from each other.
ANOVA checks the impact of one or more factors by comparing the means of different samples.

We can use ANOVA to prove or disprove if all the medication treatments were equally effective or not.

Nidhi Singh