**STATISTICS WORKSHEET 4**

**Q1to Q15 are descriptive types. Answer in brief.**

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

ANS1. The Central Limit Theorem tells us that, as the sample sizes increase, the sampling distribution of the mean will normally be distributed even though the data within each sample is not normally distributed.

IMPORTANCE:

1. it allows us to safely assume that the sampling distribution of the mean will be normal in most cases
2. we can take advantage of statistical techniques that assume a normal distribution
3. It gives us a certain distribution over our estimations. We can utilize this to pose an inquiry about the probability of an estimate that we make

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

Sampling is a method used for statistical analysis in which a predetermined number of findings are taken from a larger population. The technique used to survey a broader population depends on the type of study being conducted, but can involve simple random sampling or systematic sampling.

Broadly there are two types of sampling

1. Probability
2. Non Probability

Probability sampling

* Simple random
* Stratified
* Cluster
* Systematic

Non Probability

* Quota sampling
* Convenience sampling
* Snowball sampling

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

ANS3. A type I error (false-positive) occurs if we reject a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the we fail to reject a null hypothesis that is actually false in the population.

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

Ans4. The normal distribution is a probability function that describes how the variable values are distributed. It is a symmetrical distribution, where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper are equal in both directions.

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

Ans5. Correlation is a measure used to represent how strongly two random variables are related to each other**.** It measures the strengthand the direction of variable relationship

Covariance is a measure to indicate the extent to which two random variables change in tandem. It indicates the direction of the linear relationship between the variables

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

Univariate summarises only a single variable or feature at a time

Bivariate analysis analyses 2 variables

And Multivariate analysis is used to compare more than 2 variables at the same time

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

A sensitivity analysis indicates how different values of an independent variable affect a dependent variable under a given set of hypothesis. Sensitivity analysis helps us in forecasting using historical, true data.

Sensitivity= True Positives/ Total Positives\*100

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

ANS8.Hypothesis testing is the process where we test the assumptions regarding a population parameter. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data

H0= Null Hypothesis- The assumption we try to prove wrong

H1- Alternate Hypothesis- The assumption we are trying to prove right by proving the null hypothesis wrong

* H0: the insulin rate of patients receiving a placebo is equal to the insulin rate of patients receiving a medication.
* Ha: the insulin rate of patients receiving a placebo is different from the insulin rate of patients receiving a medication.

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

Quantitative data is the data about quantities, and therefore numbers and which can be directly measured

Qualitative data is descriptive, and regards phenomenon which can be observed but not measured, such as language. Often tools like NLP are used to convert descriptive data into quantitative format

**10. How to calculate range and interquartile range?**

The range is calculated by subtracting the lowest value from the highest value. While a large range means high variability, a small range means low variability in a distribution.

Interquartile range

Order the data from least to greatest

Find the median

Calculate the median of both the lower and upper half of the data.

The IQR is the difference between the upper and lower medians

**11. What do you understand by bell curve distribution ?**

The bell curve is a common type of distribution of the variable, also known as the normal distribution. The term "bell curve" is derived from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.

**12. Mention one method to find outliers.**

Zscore is another method to find out and remove outliers

Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

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

The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H0) of a study question is true – the definition of ‘extreme’ depends on how the hypothesis is being tested. P is also described in terms of rejecting H0 when it is actually true, however, it is not a direct probability of this state.

**14. What is the Binomial Probability Formula?**

Binomial probability refers to the probability that a [binomial experiment](https://stattrek.com/statistics/dictionary.aspx?definition=Binomial_experiment)results in exactly *x* successes.

Suppose a binomial experiment consists of n trials and results in x successes. If the probability of success on an individual trial is p, then the binomial probability is:

b(x; n, p) = nCx \* px \* qn – x

**15. Explain ANOVA and it’s applications.**

ANS15. Analysis of variance (ANOVA) is a statistical technique that is 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.

Applications

Medical diagnosis

Analysis exam performance