STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

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

ANS- The central limit theorem states that if we have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample mean is asymptotically normal.

Importance of Central Limit Theorem:
This is useful since the researcher never knows which mean in the sampling distribution corresponds to the population mean, but by taking numerous random samples from a population, the sample means will cluster together, allowing the researcher to obtain a very accurate estimate of the population mean.

2. What is sampling? How many sampling methods do you know?
ANS- Sampling means selecting the group that you will actually collect data from in your research.

There are two primary types of sampling methods that you can use in your research:

- 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 or other criteria, allowing you to easily collect data.

3. What is the difference between type1 and typeII error?
ANS- Type I error: "rejecting the null hypothesis when it is true". Type II error: "failing to reject the null hypothesis when it is false".

4. What do you understand by the term Normal distribution?
ANS- normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

5. What is correlation and covariance in statistics?
ANS- in statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other.

Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related.

6. Differentiate between univariate ,Biavariate,and multivariate analysis.
   - ANS- **Univariate**- One variable is analyzed at a time. Objective is to describe the variable. Example- How many students are graduating with "Analytics" degree?
   - **Bivariate**- Two variables are analyzed together for any possible association or empirical relationship. Example- What is the correlation between "Gender" and graduation with "Analytics" degree?

- **Multivariate**- More than two variables are analyzed together for any possible association or interactions. Example – What is correlation between "Gender", "Country of Residence" and graduation with " Analytics" degree? Any statistical modeling exercise such as Regression, Decision Tree, SVM, Clustering are multivariate in nature.

7. What do you understand by sensitivity and how would you calculate it?
   ANS- The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

   1. Define the base case of the model;
   2. Calculate the output variable for a new input variable, leaving all other assumptions unchanged;
   3. Calculate the sensitivity by dividing the % change in the output variable over the % change in the input variable.

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

   Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution.

   In hypothesis testing there are two mutually exclusive hypotheses; the Null Hypothesis (H0) and the Alternative Hypothesis (H1). One of these is the claim to be tested and based on the sampling results the claim will either be supported or not.
   Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. In our case it will be the average height of students in the batch is 100. Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed.

9. What is quantitative data and qualitative data?
   ANS- Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

10. How to calculate range and interquartile range?

ANS- 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.
R=H-L
To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

11. What do you understand by bell curve distribution ?
    ANS- A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.

12. Mention one method to find outliers.

ANS- Data Visualization Method - You can use software to visualize your data with a box plot, or a box-and-whisker plot, so you can see the data distribution at a glance. This type of chart highlights minimum and maximum values (the range), the median, and the interquartile range for your data.

Many computer programs highlight an outlier on a chart with an asterisk, and these will lie outside the bounds of the graph.

13. What is p-value in hypothesis testing?
    ANS- The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.

14. What is the Binomial Probability Formula?
    ANS- $P(x) = {}^{n}C_{x} \cdot p^{x}(1-p)^{n-x}$

15. Explain ANOVA and it's applications.

ANS- ANOVA is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

Applications - It can be in different fields of sciences, i.e. all the problems of testing more than three groups. ANOVA is used in a business context to help manage budgets by comparing your budget to costs to help manage revenue and inventory, for example. ANOVA can also be used to forecast trends by analyzing patterns in data to better understand the future performance of sales.