STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

1. What is central limit theorem and why is it important?
2. What is sampling? How many sampling methods do you know?
3. What is the difference between type1 and typeII error?
4. What do you understand by the term Normal distribution?
5. What is correlation and covariance in statistics?
6. Differentiate between univariate ,Biavariate,and multivariate analysis.
7. What do you understand by sensitivity and how would you calculate it?
8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
9. What is quantitative data and qualitative data?
10. How to calculate range and interquartile range?
11. What do you understand by bell curve distribution ?
12. Mention one method to find outliers.
13. What is p-value in hypothesis testing?
14. What is the Binomial Probability Formula?
15. Explain ANOVA and it's applications.

1.A) The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

The standard approach will be to calculate the average simply:

- Calculate the total marks of all the students in Class X
- Add all the marks
- Divide the total marks by the total number of students

2.A) 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
Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. The method of sampling depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

3.A) A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

As you analyze your own data and test hypotheses, understanding the difference between Type I and Type II errors is extremely important, because there's a risk of making each type of error in every analysis, and the amount of risk is in your control.

Type 1 error is more dangerous than Type 2 error because you are convicting the innocent person. But if you can see then Type 2 error is also dangerous because freeing a guilty can bring more chaos in societies because now the guilty can do more harm to society.

4.A) A normal distribution is **a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme**. The middle of the range is also known as the mean of the distribution.

The normal distribution is also known as a *Gaussian distribution* or *probability bell curve*. It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean.

5.A) Correlation and Covariance both measure only the linear relationships between two variables. This means that when the correlation coefficient is zero, the covariance is also zero. Both correlation and covariance measures are also unaffected by the change in location.
However, when it comes to making a choice between covariance vs correlation to measure relationship between variables,

6.A) **Univariate data** –
This type of data consists of **only one variable**. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

Bivariate data –
This type of data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.Example of bivariate data can be temperature and ice cream sales in summer season.

7.A) The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

What is sensitivity and its formula?
Assuming this function is is n= f(A, lmbda) , Then the amplitude sensitivity S at a given lambda= (dn/n) /(dA/A), one can calculate S at different lambdas and plot S versus lambda.
Sensitivity analysis is used to identify how much variations in the input values for a given variable impact the results for a mathematical model.
Sensitivity analysis can identify the best data to be collected for analyses to evaluate a project's return on investment (ROI).

8.A) What is hypothesis testing?
The procedure to decide whether the sample data are agreeable or consistent with the null hypothesis is called statistical hypothesis testing or simply hypothesis testing or test of significance.

What is H0 and H1?
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 (which infers a similar measurement in the population), the claim will either be supported or not.

What is H0 and H1 for two-tail test?
Our null hypothesis is that the mean is equal to x. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x.

9.A) What is quantitative data and qualitative data?

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.A) The IQR describes the middle 50% of values when ordered from lowest to highest. 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.A) 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.
A bell curve is a common type of distribution for a variable, also known as the normal distribution. The term "bell curve" originates from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.

12.A) Mention one method to find outliers?

Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your statistical analyses and skew the results of any hypothesis tests.

It’s important to carefully identify potential outliers in your dataset and deal with them in an appropriate manner for accurate results.

There are four ways to identify outliers:

Sorting method
Data visualization method
Statistical tests (z scores)
Interquartile range method

Sorting method
You can **sort** quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.
This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

13.A) What is p-value in hypothesis testing?

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
The smaller the *p* value, the more likely you are to reject the null hypothesis.

14.A) What is the Binomial Probability Formula?

Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment). If the probability of success on an individual trial is p , then the binomial probability is $nCx \cdot px \cdot (1-p)n-x$.

15.A) Explain ANOVA and it's applications.?

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.

Analysis of Variance (ANOVA) is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.

The Anova test is performed by comparing two types of variation, the variation between the sample means, as well as the variation within each of the samples. The below mentioned formula represents one-way Anova test statistics: Alternatively, F = MST/MSE. MST = SST/ p-1.