WORKSHEET - 4 STATISTICS

Q1to Q15 are descriptive types. Answer in brief.

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

Answer - The Central Limit Theorem states that even if the data inside each sample are not normally distributed, as sample numbers increase, the sampling distribution of the mean will become normally distributed.

Statistics benefits from the Central Limit Theorem because it makes it safe to assume that the sampling distribution of the mean will be typically normal. As a result, we can benefit from statistical methods that presume a normal distribution.

2) What is sampling? How many sampling methods do you know?

Answer - Sampling is a technique for choosing certain individuals or a small portion of the population in order to draw conclusions about the population as a whole and estimate its characteristics. Researchers frequently utilise various sampling techniques in market research so they do not have to study the full community in order to gather useful information. It is also a time- and money-efficient method, serving as the cornerstone of every research design. Software for research surveys can employ sampling strategies for the best derivation.

There are two types of sampling methods:-

- Probability sampling: Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.

- Non-probability sampling: In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

3) What is the difference between type I and type II error?

Answer -

Basis for comparison	Type I error	Type II error
Definition	Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.	Type II error is the error that occurs when the null hypothesis is accepted when it is not true.
Also termed	Type I error is equivalent to false positive.	Type II error is equivalent to a false negative.
Meaning	It is a false rejection of a true hypothesis.	It is the false acceptance of an incorrect hypothesis.
Symbol	Type I error is denoted by α.	Type II error is denoted by β.
Probability	The probability of type I error is equal to the level of significance.	The probability of type II error is equal to one minus the power of the test.
Reduced	It can be reduced by decreasing the level of significance.	It can be reduced by increasing the level of significance.

Cause	It is caused by luck or chance.	It is caused by a smaller sample size or a less powerful test.
What is it?	Type I error is similar to a false hit.	Type II error is similar to a miss.
Hypothesis	Type I error is associated with rejecting the null hypothesis.	Type II error is associated with rejecting the alternative hypothesis.
When does it happen?	It happens when the acceptance levels are set too lenient.	It happens when the acceptance levels are set too stringent.

4) What do you understand by the term Normal distribution?

Answer - Normal distribution, also known as the 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. In graph form, normal distribution will appear as a bell curve.

- A normal distribution is the proper term for a probability bell curve.
- In a normal distribution, the mean is zero and the standard deviation is 1. It has zero skew and kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- In reality, most pricing distributions are not perfectly normal.

5) What is correlation and covariance in statistics?

Answer - Given that variance describes the range of a random variable, covariance describes the interaction between two random variables. Covariance can be either negative or positive, in contrast to Variance, which is nonnegative (or zero, of course). Two random variables have a tendency to vary in the same direction when the covariance value is positive. On the other hand, when the covariance value is negative, the random variables tend to vary in the opposite way. Since they have no influence on one another and so do not vary together, two random variables that are independent have a covariance of 0, which makes sense (although this relationship may not hold in the other way).For two random variables X and Y, you can define the Covariance Cov(X, Y) as:

$\mathrm{Cov}(X, Y) = E\left((X-E(X))\ (X-E(Y))\right)$

Correlation is the Covariance divided by the standard deviations of the two random variables. Of course, you could solve for Covariance in terms of the Correlation; we would just have the Correlation times the product of the Standard Deviations of the two random variables. Consider the Correlation of a random variable with a constant. We know, by definition, that a constant has 0 variance our mathematical definition is as follows for random variables XX and YY:

$\rho=\mathrm{Corr}(X, Y) =\mathrm{Cov}(X, Y)/\sigma(x)\ \sigma(y)$

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

Answer - The simplest type of data analysis, known as a univariate analysis, involves only one variable being present in the data being examined. Being a single variable, it doesn't deal with relationships or causation. The primary goals of univariate analysis are to explain the data and identify any patterns that may be present. Only one variable is analysed at a time in univariate statistics. Two variables are compared in bivariate statistics. Multiple variables are compared using multivariate statistics..

Finding a relationship between two different variables is done via bivariate analysis. You can occasionally get a sense of what the data is trying to tell you by doing something as easy as making a scatterplot by contrasting one variable against another on a Cartesian plane (imagine an X and Y axis). There is a relationship or correlation between the two variables if the data looks to fit a line or curve. Consider plotting calorie consumption versus weight as one example.

The analysis of three or more variables is referred to as multivariate analysis. Depending on your objectives, multivariate analysis can be done in a variety of methods.

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

Answer - According to a specific set of assumptions, sensitivity analysis evaluates how various values of an independent variable impact a specific dependent variable. In other words, sensitivity analyses look at how different types of uncertainty in a mathematical model affect the overall level of uncertainty in the model.

Below are mentioned the steps used to conduct sensitivity analysis:

Firstly the base case output is defined; say the NPV at a particular base case input value (V1) for which the sensitivity is to be measured. All the other inputs of the model are kept constant. Then the value of the output at a new value of the input (V2) while keeping other inputs constant is calculated.

Find the percentage change in the output and the percentage change in the input. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

This process of testing sensitivity for another input (say cash flows growth rate) while keeping the rest of inputs constant is repeated until the sensitivity figure for each of the inputs is obtained. The conclusion would be that the higher the sensitivity figure, the more sensitive the output is to any change in that input and vice versa.

8) What is hypothesis testing? What are H0 and H1? What is H0 and H1 for a two-tail test?

Answer - Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.

Alternative Hypothesis: H1: The hypothesis that we are interested in proving.

Null hypothesis: H0: The complement of the alternative hypothesis.

The default null hypothesis for a 2-sample t-test is that the two groups are equal. You can see in the equation that when the two groups are equal, the difference (and the entire ratio) also equals zero.

9) What is quantitative data and qualitative data?

Answer - Quantitative data is defined as the value of data expressed as counts or numbers, where each item of data is given a specific numerical value. This data is any quantifiable information that can be utilised for statistical analysis and mathematical computations so that judgments in the actual world can be based on the results of these calculations. To provide answers to questions like "How many?" "How often?" and "How much?" quantitative data is used. Mathematical approaches can be used to conveniently evaluate and verify this data.

The definition of qualitative data is information that approximates and characterises. It is possible to notice and document qualitative data. The nature of this data type is not numerical. Focus groups, one-on-one interviews, observations, and other similar techniques are used to gather this kind of data. In statistics, categorical data, or information that can be categorised based on the characteristics and traits of an object or phenomena.

10) How to calculate range and interquartile range?

Answer - Range = highest-lowest

Interquartile Range Formula for a given set of data can be expressed as: IQR =

Q3 - Q1 where,

IQR = Interquartile range
Q1 = First Quartile

Q3 = Third Quartile

While the range gives you the spread of the whole data set, the interquartile range gives you the spread of the middle half of a data set.

The range of a dataset is the difference between the largest and smallest values in that dataset. For example, in the two datasets below, dataset 1 has a range of 20 – 38 = 18 while dataset 2 has a range of 11 – 52 = 41. Dataset 2 has a broader range and, hence, more variability than dataset 1. The interquartile range is the middle half of the data. To visualize it, think about the median value that splits the dataset in half. Similarly, you can divide the data into quarters. Statisticians refer to these quarters as quartiles and denote them from low to high as Q1, Q2, and Q3. The lowest quartile (Q1) contains the quarter of the dataset with the smallest values. The upper quartile (Q4) contains the quarter of the dataset with the highest values. The interquartile range is the middle half of the data that is in between the upper and lower quartiles. In other words, the interquartile range includes the 50% of data points that fall between Q1 and Q3

WORKSHEET-4

11) What do you understand by bell curve distribution?

Answer - The graphical representation of a normal probability distribution, whose underlying standard deviations from the mean form the curved bell shape, is referred to as having a "bell curve." The variability of data dispersion in a set of provided values around the mean is measured using a standard deviation.

12) Mention one method to find outliers.

Answer - Boxplots display asterisks or other symbols on the graph to indicate explicitly when datasets contain outliers. These graphs use the interquartile method with fences to find outliers. All data points beyond the IQR limit are considered outliers.

13) What is p-value in hypothesis testing?

Answer - In statistical hypothesis testing, the p-value or probability value is, for a given statistical model, the probability that, when the null hypothesis is true, the statistical summary (such as the absolute value of the sample mean the difference between two compared groups) would be greater than or equal to the actual observed results

14) What is the Binomial Probability Formula?

Answer - The Binomial Probability distribution of exactly x successes from n number of trials is given by the below formula:

$P (X) = nCx px qn – x$

Where, n = Total number of trials x = Total number of successful trials p = probability of success in a single trial q = probability of failure in a single trial = 1-p

15) Explain ANOVA and it's applications.

Answer - ANOVA is a statistical method used to determine whether the means of two or more groups differ from one another significantly. ANOVA compares the means of various samples to examine the influence of one or more factors.

Example:-

To choose the best materials to employ in the construction of a product for a consumer, a manufacturing facility would probably do an ANOVA test. Which metal is the strongest to buy from may need to be tested by the company. The corporation may be seeking for ways to save costs while still producing a high-quality product if the cost of three different types of metals is significantly different from one another.