WORKSHEET - 4 STATISTICS

Q1to Q15 are descriptive types. Answer in brief.

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

Ans. The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not normally distributed.

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution.

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

Ans. Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate the characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights. It is also a time-convenient and cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in research survey software for optimum 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?

Ans.

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?

Ans. 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, normaldistribution 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?

Ans. We know that variance measures the spread of a random variable, so Covariance measures how two random variables vary together. Unlike Variance, which is non-negative, Covariance can be negative or positive (or zero, of course). A positive value of Covariance means that two random variables tend to vary in the same direction, a negative value means that they vary in opposite directions, and a 0 means that they don't vary together. If two random variables are independent, their Covariance is 0, which makes sense because they don't affect each other and thus don't vary together

(This relation doesn't necessarily hold in the opposite direction,). For two random variables X and Y, you can define the Covariance Cov(X, Y) as:

$$Cov(X, Y) = E((X - E(X))(X - E(Y)))$$

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:

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

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

Ans. Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Since it's a single variable it doesn't deal with causes or relationships. The main purpose of the univariate analysis is to describe the data and find patterns that exist within it. Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

Bivariate analysis is used to find out if there is a relationship between two different variables. Something as simple as creating a scatterplot by plotting one variable against another on a Cartesian plane (think X and Y axis) can sometimes give you a picture of what the data is trying to tell you. If the data seems to fit a line or curve then there is a relationship or correlation between the two variables. For example, one might choose to plot caloric intake versus weight.

Multivariate analysis is the analysis of three or more variables. There are many ways to perform multivariate analysis depending on your goals.

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

Ans. Sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty.

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?

Ans. 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?

Ans. Quantitative data is defined as the value of data in the form of counts or numbers where each data set has a unique numerical value associated with it. This data is any quantifiable information that can be used for mathematical calculations and statistical analysis, such that real-life decisions can be made based on these mathematical derivations. Quantitative data is used to answer questions such as “How many?”, “How often?”, “How much?”. This data can be verified and can also be conveniently evaluated using mathematical techniques.

Qualitative data is defined as the data that approximates and characterizes. Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods. Qualitative data in statistics is also known as categorical data – data that can be arranged categorically based on the attributes and properties of a thing or a phenomenon

10) How to calculate range and interquartile range?

Ans. 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

11) What do you understand by bell curve distribution?

Ans. The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the 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. 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?

Ans. 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?

Ans. 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 its applications.

Ans. 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.

Example: -

A manufacturing plant would most likely use an ANOVA test to determine the best materials to use to build a product for a customer. The company may need to test which metal is the sturdiest to buy from. If the cost of three different types of metals issignificantly different in price, the company may be looking for ways to save money but still provide a quality product.