STATISTICS ASSIGNMENT:

1) The central limit theorem relies on the concept of a **sampling distribution**, which is the probability distribution of a **statistic** for a large number of samples taken from a population.

Imagining an experiment may help you to understand sampling distributions:

- Suppose that you draw a random sample from a population and calculate a statistic for the sample, such as the mean.
- Now you draw another random sample of the same size, and again calculate the mean.
- You repeat this process many times, and end up with a large number of means, one for each sample.

The distribution of the sample means is an example of a **sampling distribution.**

The central limit theorem says that the sampling distribution of the mean will always be **normally distributed**, as long as the sample size is large enough. Regardless of whether the population has a normal, poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

A normal distribution is a symmetrical, bell-shaped distribution, with increasingly fewer observations the further from the center of the distribution.

Central limit theorem formula

Fortunately, you don't need to actually repeatedly sample a population to know the shape of the sampling distribution. The parameters of the sampling distribution of the mean are determined by the parameters of the population:

- The mean of the sampling distribution is the mean of the population.

$$\mu_{\bar{x}} = \mu$$

- The standard deviation of the sampling distribution is the standard deviation of the population divided by the square root of the sample size.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

We can describe the sampling distribution of the mean using this notation:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

2) In statistics, the **sampling method** or **sampling technique** is the process of studying the population by gathering information and analyzing that data. It is the basis of the data where this sample space enormous.

There are several different sampling techniques available, and they can be subdivided into two groups. All these methods of sampling may involve specifically targeting hard or approach to reach groups.

In Statistics, there are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are::

- Probability Sampling
- Non-probability Sampling

3)

There are primarily two types of errors that occur, while hypothesis testing is performed, i.e. either the researcher rejects H_0, when H_0 is true, or he/she accepts H_0 when in reality H_0 is false. So, the former represents **type I error** and the latter is an indicator of **type II error**.

The testing of hypothesis is a common procedure; that researcher use to prove the validity, that determines whether a specific hypothesis is correct or not. The result of testing is a cornerstone for accepting or rejecting the null hypothesis (H_0). The null hypothesis is a proposition; that does not expect any difference or effect. An alternative hypothesis

4) 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 graphical form, the normal distribution appears as a bellcurve.

KEY TAKEAWAYS

- The 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 a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- Many naturally-occurring phenomena tend to approximate the normal distribution.
- In finance, most pricing distributions are not, however, perfectly normal.

5)

Covariance –

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
3. It is used for the linear relationship between variables.
4. It gives the direction of relationship between variables.

Correlation –

1. It show whether and how strongly pairs of variables are related to each other.
2. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
3. In this variable are indirectly related to each other.
4. It gives the direction and strength of relationship between variables.

6) **1. 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.

2. 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.

3. Multivariate data –

When the data involves **three or more variables**, it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved.Some of the techniques are regression analysis,path analysis,factor analysis and multivariate analysis of variance (MANOVA).

7)

Sensitivity and specificity: Sensitivity is the percentage of persons with the disease who are correctly identified by the test. Specificity

is the percentage of persons without the disease who are correctly excluded by the test. Clinically, these concepts are important for confirming or excluding disease during screening. Ideally, a test should provide a high sensitivity and specificity.
Sensitivity = TP/(TP + FN) and Specificity = TN/(TN + FP).
Abbreviations: TP, true positive; TN, true negative; FP, false positive; FN

8) What is H0 and H1 in hypothesis?

Alternative Hypothesis: H1: The hypothesis that we are interested in proving. Null hypothesis: H0: The complement of the alternative hypothesis. Type I error: reject the null hypothesis when it is correct. It is measured by the level of significance , i.e., the probability of type I error.

What is H0 value?

The null hypothesis (H0) is a statement of no difference, no association, or no treatment effect. The alternative hypothesis, Ha is a statement of difference, association, or treatment effect. H0 is assumed to be true until proven otherwise. However, Ha is the

9) Quantitative data refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it's quantitative data.

Unlike quantitative data, qualitative data cannot be measured or counted. It's descriptive, expressed in terms of language rather than numerical values.

10) Range

In Statistics, the **range** is the smallest of all the measures of dispersion. It is the difference between the two extreme conclusions of the distribution. In other words, the range is the difference between the maximum and the minimum observation of the distribution.

It is defined by

Range = $X_{max} - X_{min}$

Where X_{max} is the largest observation and X_{min} is the smallest observation of the variable values.

Interquartile Range Formula

The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below

Interquartile range = Upper Quartile – Lower Quartile = $Q_3 - Q_1$

where Q_1 is the first quartile and Q_3 is the third quartile of the series.

11) 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.

The highest point on the curve, or the top of the bell, represents the most probable event in a series of data (its mean, mode, and *median* in this case), while all other possible occurrences are symmetrically distributed

around the mean, creating a downward-sloping curve on each side of the peak. The width of the bell curve is described by its standard deviation.

12) outilers method: Interquartile Range Formula

The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below

Interquartile range = Upper Quartile – Lower Quartile = $Q_3 - Q_1$

where Q_1 is the first quartile and Q_3 is the third quartile of the series.

13) The P-value method is used in Hypothesis Testing to check the significance of the given Null Hypothesis. Then, deciding to reject or support it is based upon the specified significance level or threshold.

A P-value is calculated in this method which is a test statistic. This statistic can give us the probability of finding a value (Sample Mean) that is as far away as the population mean.

The P in P-value stands for **probabilty**

Based on that probability and a significance level, we Reject or Fail to Reject the Null Hypothesis.

Generally, the lower the p-value, the higher the chances are for Rejecting the Null Hypothesis and vice versa.

Also, we make use of the Z-table to perform this process

- 14) The binomial probability formula for any random variable x is given by P (x : n, p) = n C x p x q n-x n = the number of trials.
- 15) Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression.

- Analysis of variance, or 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.

- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.
-