STATISTICS WORKSHEET – 4

1. CENTRAL LIMIT THEORAM :

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

IMPORTANCE OF CLT:

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. SAMPLING :

In statistics, quality assurance, and survey methodology, sampling is the selection of a subset (a statistical sample) of individuals from within a statistical population to estimate characteristics of the whole population. Statisticians attempt to collect samples that are representative of the population in question. Sampling has lower costs and faster data collection than measuring the entire population and can provide insights in cases where it is infeasible to measure an entire population.

Types of Sampling:

- **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.

BASIS FOR COMPARISON	TYPE I ERROR	TYPE II ERROR
Meaning	Type I error refers to non-acceptance of hypothesis which ought to be accepted.	Type II error is the acceptance of hypothesis which ought to be rejected.
Equivalent to	False positive	False negative
What is it?	It is incorrect rejection of true null hypothesis.	It is incorrect acceptance of false null hypothesis.
Represents	A false hit	A miss
Probability of committing error	Equals the level of significance.	Equals the power of test
Indicated by	Greek letter 'α'	Greek letter 'β'

4. <u>TERM NORMAL DISTRIBUTION :</u>

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. **Covariance** is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable. The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicates a positive relationship.

   The greater this number, the more reliant the relationship. Positive covariance denotes a direct relationship and is represented by a positive number.

   A negative number, on the other hand, denotes negative covariance, which indicates an inverse relationship between the two variables. Covariance is great for defining the type of relationship, but it's terrible for interpreting the magnitude.

   Correlation is a measure that determines the degree to which two or more random variables move in sequence. When an equivalent movement of another variable reciprocates the movement of one variable in some way or another during the study of two variables, the variables are said to be correlated.

6. 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 are done to find out the relationship among the two variables. Example of bivariate data can be temperature and ice cream sales in summer season.

- **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 :

The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as sensitive analysis. Its usage will depend on one or more input variables within the specific boundaries, such as the effect that changes in interest rates will have on a bond's price.

It is also known as the what – if analysis. Sensitivity analysis can be used for any activity or system. All from planning a family vacation with the variables in mind to the decisions at corporate levels can be done through sensitivity analysis.

Calculation of Sensitivity:

- Firstly, the analyst is required to design the basic formula, which will act as the output formula. For instance, say NPV formula can be taken as the output formula.
- Next, the analyst needs to identify which are the variables that are required to be sensitized as they are key to the output formula. In the NPV formula in excel the cost of capital and the initial investment can be the independent variables.
- Next, determine the probable range of the independent variables.
- Next, open an excel sheet and then put the range of one of the independent variable along the rows and the other set along with the columns.
- Range of 1st independent variable
- Range of 2nd independent variable

8. HYPOTHESIS TESTING :

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. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases in the following descriptions.

Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

H0: The null hypothesis, H0, is a statistical proposition stating that there is no significant difference between a hypothesized value of a population parameter and its value estimated from a sample drawn from that population.

H1: is a statistical proposition stating that there is a significant difference between a hypothesized value of a population parameter and its estimated value. When the null hypothesis is tested, a decision is either correct or incorrect.

H0 & H1 FOR TWO-TAIL TESTING:

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.

$H0 : \mu = 100$

Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed. "In our case, Tedd believes (Claims) that the actual value has changed". He doesn't know whether the average has gone up or down, but he believes that it has changed and is not 100 anymore.

$H1: \mu \neq 100$

9. **QUALITATIVE DATA** is non-statistical and is typically unstructured or semi-structured. This data isn't necessarily measured using hard numbers used to develop graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers.
Qualitative data can be used to ask the question "why." It is investigative and is often open-ended until further research is conducted. Generating this data from qualitative research is used for theorizations, interpretations, developing hypotheses, and initial understandings.

Qualitative data can be generated through:

- Texts and documents
- Audio and video recordings
- Interview transcripts and focus groups
- Observations and notes

➢ **QUANTITATIVE DATA:** Contrary to qualitative data, quantitative data is statistical and is typically structured in nature – meaning it is more rigid and defined. This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

Whereas qualitative is open for exploration, quantitative data is much more concise and close-ended. It can be used to ask the questions "how much" or "how many," followed by conclusive information.

Quantitative data can be generated through:

- Tests
- Experiments
- Surveys
- Market reports
- Metrics

10. **CALCULATION OF RANGE**: To calculate the range, you need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum). The range only takes into account these two values and ignore the data points between the two extremities of the distribution

CALCULATION OF INTERQUARTILE RANGE: 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.D

11. A BELL CURVE DISTRIBUTION :

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. Bell curves are visual representations of normal distribution, also called Gaussian distribution.

A normal distribution curve, when graphed out, typically follows a bell-shaped curve, hence the name. While the precise shape can vary according to the distribution of the population, the peak is always in the middle and the curve is always symmetrical.

Bell curves are useful for quickly visualizing a data set's mean, mode and median because when the distribution is normal, the mean, median and mode are all the same.

The long tail refers to the part of the bell curve that stretches out in either direction. If the diagram above represents a population under study, the fat area under the bell curve is where most of the population falls.

12. SORTING METHOD :

An easy way to identify outliers is to sort your data, which allows you to see any unusual data points within your information. Try sorting your data by ascending or descending order, then examine the data to find outliers. An unusually high or low piece of data could be an outlier.

For example, if you have these numbers in ascending order: 3, 6, 7, 10 and 54, you can see that 54 are a lot larger than the rest of the data points. Statisticians would consider 54 an outlier

13. P-VALUE IN HYPOTHESIS TESTING :

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming

that the null hypothesis is correct. The p-value serves as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.
P-value is often used to promote credibility for studies or reports by government agencies. For example, the U.S. Census Bureau stipulates that any analysis with a p-value greater than 0.10 must be accompanied by a statement that the difference is not statistically different from zero. The Census Bureau also has standards in place stipulating which p-values are acceptable for various publications.

14. BINOMIAL PROBABILITY FORMULA :

The binomial distribution formula: P(X= x) = nCxpxqn-x, where = 0, 1, 2, 3 … P(X = 6) = 105/512. Hence, the probability of getting exactly 6 heads is 105/512.

15. ANOVA :

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 study.

$$F = \frac{MSE}{MST}$$

Where:

F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

APPLICATIONS OF ANOVA:

1. ANOVA is designed to detect differences among means from populations subject to different treatments.
2. ANOVA is a joint test
   •The equality of several populations' means is tested simultaneously or jointly.
3. ANOVA tests for the equality of several population means by looking at two estimators of the population variance (hence analysis of variance)