STATICAL WORKSHEET = 4

1. The central limit theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size gets larger no matter what the shape of the population distribution, The central limit theorem is important because it is used in hypothesis testing and also to calculate confidence intervals.

2. Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

   Five basic sampling methods

   =>Simple random

   =>Convenience

   =>Systematic

   =>Cluster

   =>Stratified

3. The points given below are substantial so far as the differences between type I and type II error is concerned:

=> Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false.

=> Type I error or otherwise known as false positives, in essence, the positive result is equivalent to the refusal of the null hypothesis. In contrast, Type II error is also known as false negatives, i.e. negative result, leads to the acceptance of the null hypothesis.

=> When the null hypothesis is true but mistakenly rejected, it is type I error. As against this, when the null hypothesis is false but erroneously accepted, it is type II error.

=> Type I error tends to assert something that is not really present, i.e. it is a false hit. On the contrary, type II error fails in identifying something, that is present, i.e. it is a miss.

=> The probability of committing type I error is the sample as the level of significance. Conversely, the likelihood of committing type II error is same as the power of the test.

=> Greek letter ‘α’ indicates type I error. Unlike, type II error which is denoted by Greek letter ‘β’.

4. The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the centre is a mirror image of the left side. The area under the normal distribution curve represents probability and the total area under the curve sums to one. Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon.

5. Covariance = Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).The values of covariance can be any number between the two opposite infinities. Also, it’s important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.

Correlation = Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables. It not only shows the kind of relation (in terms of direction) but also how strong the relationship is. Thus, we can say the correlation values have standardized notions, whereas the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance. It can assume values from -1 to +1. To determine whether the covariance of the two variables is large or small, we need to assess it relative to the standard deviations of the two variables. To do so we have to normalize the covariance by dividing it with the product of the standard deviations of the two variables, thus providing a correlation between the two variables.

6. **Univariate** = Univariate analysis is the simplest form of data analysis where the data being analysed contains only one variable. Since it's a single variable it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it. You can think of the variable as a category that your data falls into. One example of a variable in univariate analysis might be "age". Another might be "height". Univariate analysis would not look at these two variables at the same time, nor would it look at the relationship between them.Some ways you can describe patterns found in univariate data include looking at mean, mode, median, range, variance, maximum, minimum, quartiles, and standard deviation. Additionally, some ways you may display univariate data include frequency distribution tables, bar charts, histograms, frequency polygons, and pie charts.

Bivariate = 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 = Multivariate analysis is the analysis of three or more variables. There are many ways to perform multivariate analysis depending on your goals. Some of these methods include:

- Additive Tree
- Canonical Correlation Analysis
- Cluster Analysis
- Correspondence Analysis / Multiple Correspondence Analysis
- Factor Analysis
- Generalized Procrustean Analysis
- MANOVA
- Multidimensional Scaling
- Multiple Regression Analysis
- Partial Least Square Regression
- Principal Component Analysis / Regression / PARAFAC
- Redundancy Analysis.

7. Sensitivity analysis is an analysis technique that works on the basis of what-if analysis like how independent factors can affect the dependent factor and is used to predict the outcome when analysis is performed under certain conditions. It is commonly used by investors who takes into consideration the conditions that affect their potential investment to test, predict and evaluate result.The formula for sensitivity analysis is basically a financial model in excel where the analyst is required to identify the key variables for the output formula and then assess the output based on different combinations of the independent variables. Mathematically, the dependent output formula is represented as,
   $Z = X^2 + Y^2$

8.

Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.

There are 5 main steps in hypothesis testing:

- ⇨ State your research hypothesis as a null hypothesis and alternate hypothesis (H_o) and (Ha or H1).
- ⇨ Collect data in a way designed to test the hypothesis.
- ⇨ Perform an appropriate statistical test.

- ⇨ Decide whether to reject or fail to reject your null hypothesis.
- ⇨ Present the findings in your results and discussion section.

9.

Quantitate data = 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. Quantitative data can tell you "how many," "how much," or "how often"—for example, how many people attended last week's webinar? How much revenue did the company make in 2019? How often does a certain customer group use online banking? To analyse and make sense of quantitative data, you'll conduct statistical analyses.

Qualitative data = Unlike quantitative data, qualitative data cannot be measured or counted. It's descriptive, expressed in terms of language rather than numerical values. Researchers will often turn to qualitative data to answer "Why?" or "How?" questions. For example, if your quantitative data tells you that a certain website visitor abandoned their shopping cart three times in one week, you'd probably want to investigate why—and this might involve collecting some form of qualitative data from the user. Perhaps you want to know how a user feels about a particular product; again, qualitative data can provide such insights. In this case, you're not just looking at numbers; you're asking the user to tell you, using language, why they did something or how they feel. Qualitative data also refers to the words or labels used to describe certain characteristics or traits—for example, describing the sky as blue or labelling a particular ice cream flavour as vanilla.

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.

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

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

IQR = 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.

The below figure shows the occurrence of median and interquartile range for the data set.

11. To be technical, the kinds of bell curves that we care about the most in statistics are actually called normal probability distributions. For what follows we'll just assume the bell curves we're talking about are normal probability distributions.

Despite the name “bell curve,” these curves are not defined by their shape. Instead, an intimidating looking formula is used as the formal definition for bell curves.
But we really don’t need to worry too much about the formula. The only two numbers that we care about in it are the mean and standard deviation. The bell curve for a given set of data has the center located at the mean. This is where the highest point of the curve or “top of the bell“ is located. A data set‘s standard deviation determines how spread out our bell curve is. The larger the standard deviation, the more spread out the curve.

12. Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean.

13. Usually, we get Sample Datasets to work on and perform data analysis and visualization and find insights. Through that analysis, we make inferences on the whole population. The population is the entire dataset, whereas Sample Datasets are chunks from the Population Dataset. Similar to population, these sample datasets also have some mean and standard deviation associated. These samples mean values usually vary from the population mean by different ranges. These ranges can deviate from very close to very far from the population mean. For making correct inferences, our sample datasets need to resemble the properties of the population dataset. So, these samples and the sample means are validated through various hypothesis testing methods like the P-value method, Critical Value method, T-test, ANOVA test, etc.

14. The binomial probability formula can be used to calculate the probability of success for binomial distributions. Binomial probability distribution along with normal probability distribution are the two probability distribution types. To recall, the binomial distribution is a type of distribution in statistics that has two possible outcomes. For instance, if you toss a coin and there are only two possible outcomes: heads or tails. In the same way, taking a test could have two possible outcomes: pass or fail.

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 study.

The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method.12 ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests. The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers."3 It was employed in experimental psychology and later expanded to subjects that were more complex.