STATISTICS WORKSHEET-4

1. The Central Limit Theorem (CLT) is a statistical theory that posits that the mean and standard deviation derived from a sample, will accurately approximate the mean and standard deviation of the population the sample was taken from as the size of the sample increases. 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 is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual. There are different types of sampling techniques available:
   (1) Probability Sampling:
      i. Simple Random Sampling
      ii. Systematic Samling
      iii. Stratified Sampling
      iv. Cluster

   (2) Non-Probability Sampling:
      v. Convenience Sampling
      vi. Quota
      vii. Judgement Sampling
      viii. Snowball Sampling

3. There are primarily two types of errors that occur, while hypothesis testing is performed, i.e. either the researcher rejects H0, when H0 is true, or he/she accepts H0 when in reality H0 is false. So, the former represents type I error and the latter is an indicator of type II error. 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.

4. Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

5. Covariance and correlation are two terms that are exactly opposite to each other. However, they both are used in statistics and regression analysis. Covariance shows us how the two variables vary, whereas correlation shows us the relationship and how they are related. Correlation and covariance are two statistical concepts used to determine the relationship between two random variables. Correlation defines how a change in one variable will impact the other, while covariance defines how two items vary together. Confusing? Let us dive in further to understand the difference between these closely related terms.

6. Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

7. 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.

8. Hypothesis testing is a statistical test based on two hypothesis: the null hypothesis (H0), and the alternative hypothesis (H1). Null Hypothesis (H0): H0 always assume there is no significant effect/difference within the specified population. Alternative Hypothesis (H1): H1 always has opposite opinion with H0. H 1: $\mu < \mu$ 0, where a decrease is hypothesized and this is called a lower-tailed test; or H 1: $\mu \neq \mu$ 0, where a difference is hypothesized and this is called a two-tailed test.

9. Quantitative data is numbers-based, countable, or measurable. Qualitative data is interpretation-based, descriptive, and relating to language. Quantitative data tells us how many, how much, or how often in calculations.

10. 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 = Xmax– Xmin.** Where Xmax is the largest observation and Xmin is the smallest observation of the variable values.

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. Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your statistical analyses and skew the results of any hypothesis tests. It's important to carefully identify potential outliers in your dataset and deal with them in an appropriate manner for accurate results. There are four ways to identify outliers one of them is Zscores method.

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.

14. The binomial distribution formula calculates the probability of getting x successes in the n trials of the independent binomial experiment. The probability is derived by a combination of the number of trials. First, the number of successes is represented by nCx. Then, it is multiplied by the probability of the success raised to the power of the number of successes. It is represented by px. Further, it is multiplied by the probability of the failure raised to the power of the difference between the number of successes and the number of trials represented by (1-p) n-x. The probability of obtaining x successes in n independent trials of a binomial experiment is given by the following formula of binomial distribution: P(X) = nCx px(1-p)n-x.

15. 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.