STATISTICS_WORKSHEET_4

1. The Central Limit Theorem states that the distribution of a sample mean that approximates the normal distribution, as the sample size becomes larger, assuming that all the samples are similar, and no matter what the shape of the population distribution.
   The central limit theorem tells us that no matter what the distribution of the population is, the shape of the sampling distribution will approach normality as the sample size increases.
2. Sampling is the research strategy of collecting data from a part of a population with a view to drawing inferences about the whole. The “population” in this sense is often termed the “universe”. Some of the sampling methods are Simple random sampling,Systematic sampling, Stratified sampling,Cluster sampling.
3. 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.
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 "bell curve".
5. Covariance is a measure to indicate the extent to which two random variables change in tandem. Correlation is a measure used to represent how strongly two random variables are related to each other. Covariance is nothing but a measure of correlation. Correlation refers to the scaled form of covariance.
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 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.
   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.
7. It is simply the term Sensitivity, we calculate it when we have to analyse the confusion metrics in classification. Sensitivity stands for true positive ratio and formula for the one is: TP/(TP+FN)
8. Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. H0 is Null Hypothesis and H1 is Alternative Hypothesis.

Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed.

9. Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables. Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.
10. 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. The IQR describes the middle 50% of values when ordered from lowest to highest. 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.
11. Bell Curve Distribution is nothing but the normal distribution. The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center 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.
12. Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. 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.
13. The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.
14. Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment). P(x) = P^x. (1-P)^1-x = P
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.