STATISTICS WORKSHEET-4

1. Central Limit Theorem says that with a large sample size, sample means are normally distributed. Normally distributed means that a group of numbers follows a bell-shaped curve. Most of the numbers cluster in the middle around the average, and there are fewer numbers at the extremes to the right and left.
   The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.
2. Sampling methods are the ways to choose people from the population to be considered in a sample survey. Samples can be divided based on following criteria.
   1) Probability samples - In such samples, each population element has a known probability or chance of being chosen for the sample.
   2) Non-probability samples - In such samples, one cannot be assured of having known probability of each population element.
3. Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true. Type II error is the error that occurs when the null hypothesis is accepted when it is not true
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.
5. Correlation is considered the best technique for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related. Given two random variables, it is the covariance between both divided by the product of the two standard deviations of the single variables, hence always between -1 and 1.
   Covariance is a measure that indicates the extent to which two random variables change in cycle. It explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.
6. Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it. Example: height of students.
   Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables. Example: temperature and ice cream sales in the summer season
   Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable. Example: data for house price prediction
7. Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.). It can be calculated as: TN/(TN+FP).
8. A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
   Ex : you have a coin and you don’t know whether the coin is fair or not. First, decide null and alternate hypothesis and alpha=0.05.
   H0 : that is a fair coin.
   H1 : the coin is not fair
9. Quantitative data are anything that can be expressed as a number, or quantified. Examples of quantitative data are scores on achievement tests, number of hours of study, or weight of a subject. Qualitative data cannot be expressed as a number. Data that represent nominal scales such as gender, social, economic status, and religious preference are usually considered to be qualitative data.
10. The interquartile range is a measure of where the middle is in a data set. Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie. That’s why it’s preferred over many other measures of spread when reporting things like school performance or SAT scores.
    The interquartile range formula is the first quartile subtracted from the third quartile:
    $$IQR = Q_3 – Q_1.$$
11. In a bell curve, the centre contains the greatest number of a value and, therefore, it is the highest point on the arc of the line. This point is referred to the mean, but in simple terms, it is the highest number of occurrences of an element.

12. Outlier values can be identified by using boxplot or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for large number of outliers the values can be substituted with either the 99th or the 1st percentile values. All extreme values are not outlier values. The most common ways to treat outlier values are:
    a. To change the value and bring in within a range
    b. To just remove the value.
13. *p*-value is defined as the probability that the data would be at least as extreme as those observed, if the null hypothesis were true. The *p*-value reflects the strength of evidence against the null hypothesis.
14. The binomial distribution consists of the probabilities of each of the possible numbers of successes in N trials for independent events that each have a probability of π. Here, the possible outcomes are two.
15. Analysis of Variance (ANOVA) is a technique which is used to compare the means of multiple samples. Whether there is a significant difference between the mean of 2 samples, can be evaluated using z-test or t-test but in case of more than 2 samples, t-test cannot be applied as it accumulates the error and it will be more difficult as the number of sample will increase (for example: for 4 samples — 12 t-test will have to be performed). The ANOVA technique enables us to perform this simultaneous test.