STATISTICS WORKSHEET-4

1. Central limit theorem says that when number of samples is significantly large, mean of the samples will be roughly equal to the mean of the population from which the sample set was taken.
   This theorem has been useful for taking various polls be it online or offline where a large group of sample represents entire population, useful for political analysis and predicting exit poll victories.

2. Sampling is a method of picking a number of data from a larger set of data, in such way that the picked dataset i.e. sample now represents the entire dataset.
   Theres various sampling methods –
   Probability sampling:
   a. Simple random sampling
   b. Systematic Sampling
   c. Stratified Sampling
   d. Cluster sampling

   Non Probability sampling:

   a. Convenience sampling
   b. Voluntary response sampling
   c. Purposive sampling
   d. Snowball sampling

3. Type 1 error applies where Null Hypothesis is true but gets rejected because of the results of statistical analysis (False Negative),
   Type 2 error applies where Null Hypothesis is actually False but we fail to reject Null Hypothesis because of the results of statistical analysis (False Positive)

4. In normal distribution, most of data occurs more around 'mean' of the data set, while the mean is set at ZERO, and standard deviation is 1. It appears in bell curve shape. As per central limit theorem, all datasets will appear normal as number of samples goes higher and higher. It is symmetrical distribution.

5. Covariance: A systematic relationship between a pair of random variables wherein a change in one variable reciprocated by an equivalent change another variable.
   Correlation: It measures strength of relationship of change in one variable due to change in another variable.
   Correlation tells the strength of the relationship or how strongly two variables relate to each other.

6. Univariate Analysis: This datatype consists of only one variable. It deals with only 1 quantity and does not analyse cause or relation of the data variation.
   Bivariate Analysis: This datatype consists of two variables. It deals with only two quantities and analyses cause and relation of the data variation with respect of 2 variables.

7. With respect to confusion matrix, Sensitivity/Recall is results, how many positives are correctly predicted.
   It can be calculated as:
   R_e=True Positive/(True Positive + False Negative)

8. Hypothesis testing is a procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is reasonable statement and should not be rejected, or an unreasonable statement and should be rejected. The hypothesis that is tested is called Null Hypothesis noted as H0, and if the statement is found unreasonable and is rejected then the research supports an Alternative Hypothesis which is noted H1 or Ha.

H0 (Null Hypothesis): A statement about the value of a population parameter that is assumed to be true for the purpose of testing.
H1 (Alternative Hypothesis): A statement about the value of a population parameter that is assumed to be true if Null Hypothesis is rejected during testing.
For TWO TAIL test:
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 refers to data that can be measured mathematically, like number of apples bought, miles run, goals scored, etc
   Qualitative data refers to data that cannot be measured mathematically, like smell, colour, taste, etc

10. Range is calculated as the arithmetic difference between highest and lowest value.
    Interquartile range is the middle 50% when data is sorted from lowest to highest. It can be calculated by below steps:
    a. Find median of lower half and upper half of the data.
    b. The arithmetic difference between median of upper half and median of lower half is Interquartile range.

11. Bell Curve distribution is a type of distribution in which the curve is at its y axis' highest point at the centre of curve and mean is at 0 for x axis. This graph is symmetrical and shaped like a bell. All other possible occurrences are symmetrically distributed around the mean, creating a downward-sloping curve on each side of the peak.

12. We can find outliers with help of boxplot.
    Find Interquartile range, then find upside outlier by adding median of upper half and (1.5 * interquartile range) then find downside outlier by subtracting median of lower half and (1.5 * interquartile range)

13. P-value is defined as the probability of observing any test statistic that is atleast as extreme as the one computed from a sample, given that the null hypothesis is True.

14. Binomial Probability formula is =

    $$P(x) = (n/x)\ p^x\ q^{(n-x)}$$

    Where,

    P = Probability

    x = 0,1,2,3…

    n = number of experiments
    p = Probability of success in a single experiment
    q = Probability of failure in a single experiment

15. ANOVA stands for ANALYSIS OF VARIANCE. Its is a statistical method used to study variance between two or more means. The inferences made on means is by analysing the variance, hence the name. It is used to test general differences in the means.
    ANOVA is helpful for testing multiple variables. It can tell you if there's significant difference in the variable means. It can be used in various ways, one of which is that it can be used to monitor your body stats like blood pressure, sugar, HB, BMI index before and after a meal, which can be used to prepare a diet plan by figuring out your metabolism rate.