WORKSHEET 4 STATISTICS

1) The central limit theorem (CLT) states that the sampling distribution of the mean approaches a normal distribution, as the sample size gets larger regardless of the population's distribution.
It is important because it allows us to assume that the sampling distribution of the mean will be normal in most cases. Eg: It can be used in population census to calculate individual income.

2) Sampling is a process of selecting a group of individuals from a population who will represent the entire population in a research. For example, if you are researching the opinion of employee on a topic, you will survey a sample of 1000 instead of getting opinion of every single employee in the industry.

There are two category of sampling
a) Probability sampling involves randomly choosing sample from the population for a research. Every individual has chance of being included in sample.The four main method in this category include

- Simple random sampling in which individual are selected randomly from population.
- Systematic sampling unlike simple random sampling involves choosing sample based on regular interval.
- Stratified sampling involves dividing population into subgroup based on characteristics like gender etc and then using simple random or systematic sampling a sample is selected for each subgroup.
- Clustering sampling involves dividing population into subgroup but each subgroup has similar characteristics to whole sample. Instead of selecting sample from each subgroup, entire subgroup is selected.

b) Non probability sampling involves non randomly choosing sample based on convenience or other criteria. Not every individual has chance of being included in sample.

- Convenience sampling involves selecting individuals that are easily accessible to researcher.
- Volunteer sampling is similar to convenience but participants volunteer instead of being selected by researcher.

3) Type I error or false positive occurs when null hypothesis is rejected when it is actually true in the population whereas type II error or false negative occurs when null hypothesis is accepted when it is actually false in population

4) A normal distribution is when the probability of distribution around the mean is higher in other words it indicates that the quantity of data points centered around the mean are higher than farther away from it.

5) Correlation indicates relationship between two set of features, showing how strongly or weakly the two features are bonded with each other. Unlike correlation, covariance measures how two variables vary from each other.

6) In univariate type of analysis, only a single variable is used to analyze the data whereas in bivariate analysis, two variables are used to analyze the data. If we need to analyze more than two variables then multivariate analysis is performed.

7) Sensitivity is used for evaluating a model's ability to predict the true positives of each available category. Sensitivity = (True Positive)/(True Positive + False Negative)

8) Hypothesis testing is a method i.e. used to decide whether the data at hand successfully supports a particular hypothesis.
H_0 or null hypothesis is statement i.e. assumed true until proven wrong based on experimental data.
H_1 or alternative hypothesis is statement that contradicts null hypothesis.

For two tailed
H0 (null hypothesis) states that the mean is equal to x
H1 (null hypothesis) states that the mean is not equal to x

9) Quantitative data are values that can be counted and are expressed in numbers whereas qualitative data are measures of 'types' and can be represented by a name, symbol, or a number code. Eg: Number of employees is quantitative data and rating of employee performance is qualitative data.

10) Range is calculated as difference between maximum value and minimum value in given set of data whereas inter quartile range is difference between 3rd quartile (75th percentile) and 1st quartile (25th percentile) of the dataset.

11) Bell curve is a graph depicting normal distribution. The top of curve indicates mean, median and mode and its standard deviation indicates the width of the bell curve around mean.

12) The inter quartile range (IQR) is the difference between Q3 (75th percentile) and Q1 (25th percentile). By multiplying this IQR 1.5 times, we can find a value which can be subtracted from Q1 to find the lower boundary and added to Q3 to find the upper boundary for the outliers. And thus, any data point that lies above upper boundary or below lower boundary are considered as the outliers.

13) P values are used in hypothesis testing to help decide whether to reject the null hypothesis. If p value <=0.05 then null hypothesis is rejected and if p value> 0.05 null hypothesis is accepted.

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
$q = 1 - p$ = the probability of getting a failure in one trial

14)

15) ANOVA tests if there is a difference in the mean somewhere in the model but it does not tell us where the difference is. It is basically used to check the means of two or more groups that are significantly different from each other. It assumes hypothesis as H0: Means of all groups are equal.