STATISTICS ASSIGNMENT

ANSWERS:

1. Central limit theorem states that the mean of a sample of data will be closer to the mean of the overall population in question, as the sample size increases, notwithstanding the actual distribution of the data. In other words, the data is accurate whether the distribution is normal or aberrant.

   Central limit theorem is important in sense that it is often used in conjunction with the law of large numbers, which states that the average of the sample means and standard deviations will come closer to equalling the population mean and standard deviation as the sample size grows, which is extremely useful in accurately predicting the characteristics of populations.

2. A **sample** is a subset of the population and the process or technique of selecting these individuals from the population to make statistical inferences and estimating the characteristics of the whole population is known as **sampling**.

   Types of sampling methods:

   Probability sampling: It is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter. There are four types of probability sampling techniques namely **simple random sampling, cluster sampling, systematic sampling and stratified random sampling**.

   Non-probability sampling: In this sampling method the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample. There are four types of non-probability sampling techniques namely **convenience sampling, judgemental/purposive sampling, snowball sampling and quota sampling**.

3. A **type I** error appears when the null hypothesis (H0) of an experiment is true, but still, it is rejected while a **type II** error appears when the null hypothesis is false but mistakenly fails to be refused. A **type I error** is also called as **false positive** and a **type II** error as **false negative**.

4. A **normal distribution** is defined by a mean (average) of zero and a standard deviation of 1.0, with a skew of zero and kurtosis = 3. In a normal distribution, approximately 68% of the data collected will fall within +/- one standard deviation of the mean; approximately 95% within +/- two standard deviations; and 99.7% within three standard deviations.

5. **Correlation** is a statistical measure that indicates how strongly two variables are related. **Covariance** is a measure of how much two random variables vary together.

6. **Univariate** analysis deals with uni i.e. only one variable. Example would be analysing students' performance based on the marks obtained in Mathematics.
   Bivariate analysis deals with bi i.e. two variable. Example would be would be analysing students' performance based on the marks obtained in Mathematics and English.
   Multivariate analysis deals with multiple variables i.e. more than two variables. Example would be analysing a students' performance based on marks obtained in Mathematics, English, Science, Social Studies.

7. **Sensitivity** (true positive rate) refers to the probability of a positive test, conditioned on truly being positive. Since it is the true positive rate it can be calculated by dividing the number of true positives with total number of positives.

   Sensitivity = TP/TP+FN

   Where;

   TP = Number of true positives

   FN = Number of false negatives

8. **Hypothesis Testing** is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

   The **Null Hypothesis** is the assumption that the event will not occur. **H0** is the symbol for it. The **Alternate Hypothesis** is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. **H1** is the symbol for it.

   In two-tail test the critical distribution area is two-sided.
   Suppose H0: mean = 50 and H1: mean not equal to 50

According to the H1, the mean can be greater than or less than 50. This is an example of a Two-tailed test.

9. The data collected on the grounds of the numerical variables are **quantitative data**. Quantitative data are more objective and conclusive in nature. The data collected on grounds of categorical variables are **qualitative data**. Qualitative data are more descriptive and conceptual in nature.

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 lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The median is considered the second quartile (Q2). The **interquartile range** is the difference between upper and lower quartiles i.e. **Q3-Q1**.

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.

12. **Outliers** can be detected using z-score method. **Z-score** tells you where the score lies on a normal distribution curve. If the z score of a data point is more than **+3** or **-3** then that data point is considered to be an outlier.

13. A **p-value** is a metric that expresses the likelihood that an observed difference could have occurred by chance. As the p-value decreases the statistical significance of the observed difference increases. If the p-value is too low, you reject the null hypothesis.

14. The **binomial distribution** is a probability distribution used in statistics that summarizes the likelihood that a value will take one of two independent values under a given set of parameters or assumptions.

    $P(x:n,p) = {}^{n}C_{x}X\ p^{x}\,(1-p)^{n-x}$

    where:

n is the number of trials (occurrences)
X is the number of successful trials
p is probability of success in a single trial
nCx is the combination of n and x. A combination is the number of ways to choose a sample of x elements from a set of n distinct objects where order does not matter and replacements are not allowed. Note that nCx=n!/(r!(n−r)!), where ! is factorial (so, 4! = 4 x 3 x 2 x 1)

15. **Analysis of variance (ANOVA)** is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

    The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.