Advanced Analytics Assignment 02

248518 Gaurav B Raskar

Q.1.

Univariate analysis: studies one variable to understand its distribution and patterns.

Bivariate analysis: examines the relationship between two variables to see how they interact.

Multivariate analysis: explores the connections between three or more variables at the same time.

Q.2.

Central Limit Theorem (CLT) states that the distribution of a sample will approximate a normal distribution (i.e., a bell curve) as the sample size becomes larger.

Q.3.

Data mining is the process of discovering patterns, trends, and useful information from large sets of data using techniques like statistics, machine learning, and database systems. Involves Data Wrangling, Data Cleaning, Data Transformation, Dimension Reduction etc.

Q.4.

Type I Error:

* Represented by α (alpha).
* Occurs when actual Null Hypothesis is False but, Measured as True.

Type II Error:

* Represented by β (beta).
* Occurs when actual Null Hypothesis is True but, Measured as False.

Q.5.

Due to Time Constraint and Cost effectiveness.

Q.6.

1) Simple Random Sampling:

Randomly select samples with given sample size.

2) Stratified Random Sampling:

Randomly selected from each Homogenous stratas/groups of population.

3) Systematic Random Sampling:

If data is extremely homogenous then, periodically selected from ordered list.

4) Cluster Sampling:

Sampling with multiple stages by selecting clusters and sub-clusters. Clusters are Heterogeneous.

Q.7.

Stratified sampling is a method where a population is divided into distinct subgroups (strata) based on specific characteristics, ensuring that each subgroup is adequately represented. Samples are then drawn from each stratum, either proportionally or equally, to create a more accurate and reliable overall sample.

Q.8.

Time series analysis is the study of data collected over time to find patterns or trends. It helps us understand how things change and can be used to predict future events, like sales or weather, based on past information.

Q.9.

Normal distribution is a bell-shaped curve that shows how data is spread around the average, which follows Gaussian Equation.

Q.10.

Outliers can be treated by identifying them using methods like the interquartile range or Z-scores, and then deciding whether to remove, transform, or analyze them separately based on their impact on the dataset.

Q.11.

CV = std./mean = 0.173/4.7 = 0.04

Q.12.

Q.13.

Normal distribution is a bell-shaped curve that shows how data is spread around the average, which follows Gaussian Equation.

Q.14.

Outliers can be treated by identifying them using methods like the interquartile range or Z-scores, and then deciding whether to remove, transform, or analyze them separately based on their impact on the dataset.

Q.15.

Linear regression is a way to understand how one thing affects another by drawing a straight line through a set of data points. It helps us predict outcomes based on the relationship between variables, like how studying hours might influence test scores.

Q.17.

ANOVA test involves comparing means to accept or reject null hypothesis. Also, known as Analysis of Variance. Which involves one or more Factor variables with two or more Factors and One Dependent Variable.

Q.18.

A hypothesis test is a statistical procedure that evaluates whether sample data provides enough evidence to support a specific claim about a population.

Q.19.

T-Test

ANOVA

Wilcox Test

U-Test

H-Test

Pearson’s Correlation

Spearman’s Correlation

Chi-Squared Test

Etc.