STATISTICS WORKSHEET

1. Which of the following is the correct formula for total variation?

ANS b) Total Variation = Residual Variation + Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.

ANS c) binomial d) none of the mentioned

3. How many outcomes are possible with Bernoulli trial?

ANS a) 2

4. If Ho is true and we reject it is called

ANS a) Type-I error

5. Level of significance is also called:

ANS d) Confidence coefficient

6. The chance of rejecting a true hypothesis decreases when sample size is:

ANS b) Increase

7. Which of the following testing is concerned with making decisions using data?

ANS b) Hypothesis

8. What is the purpose of multiple testing in statistical inference?

ANS d) All of the mentioned

9. Normalized data are centered at and have units equal to standard deviations of the original data

ANS a) 0

10. What Is Bayes' Theorem?

ANS The Bayes theorem definition (Bayes rule) is a probability measure proposed by British statistician Thomas Bayes. His findings were compiled in an essay, "*Towards Solving a Problem in the Doctrine of Chances*

The Bayes theorem determines the probability of an event A occurring based on the probability of the occurrence of event B—provided both events occur independently. The following Bayes theorem formula represents it:

$$P\ (B) = \frac{P(A)\,P(A)}{P(B)}\ or, \frac{P(A \cap B)}{P(B)}$$

- P(A|B) is the probability of event A occurring after event B.
- P(B|A) is the probability of event B occurring after event A.
- P(A) is the probability of event A occurring.
- P(B) is the probability of e vent B occurring.

11. What is z-score?

ANS Z-score is a statistical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

Z-Score Formula

The statistical formula for a value's z-score is calculated using the following formula:

$$z = (x - \mu) / \sigma$$

Where:

- z = Z-score
- x = the value being evaluated
- μ = the mean
- σ = the standard deviation

Calculating a z-score requires that you first determine the mean and standard deviation of your data. Once you have these figures, you can calculate your z-score. So, assume you have the following variables:

- x = 57
- μ = 52
- σ = 4

You would use the variables in the formula:

- z = (57 - 52) / 4
- z = 1.25

So, your selected value has a z-score that indicates it is 1.25 standard deviations from the mean.

12. What is t-test?

ANS A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times. The t-test is a test used for hypothesis testing in statistics and uses the t-statistic, the t-distribution values, and the degrees of freedom to determine statistical significance.

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

13. What is percentile?

ANS The percentile formula is used when we need to compare the exact values or numbers over the other numbers from the given data i.e. the accuracy of the number. Often percentile and percentage are taken as one but both are different concepts. A percentage is where the fraction is considered as one term while percentile is the value below the percentage found from the given data. In our day-to-day life, percentile formulas are usually helpful in finding the test scores or biometric measurements. Hence, the percentile formula is:

Percentile = (n/N) × 100

Or

The percentile of x is the ratio of the number of values below x to the total number of values multiplied by 100. i.e., the percentile formula is

Percentile = (Number of Values Below "x" / Total Number of Values) × 100

14. What is ANOVA?

ANS Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The Formula for ANOVA is:

$F=MST/MSE$

where: F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

15. How can ANOVA help

ANS A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample T-TEST. However, it results in fewer Type I Errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.

MACHINE LEARNING ASSIGNMENT – 3

1. Which of the following is an application of clustering?

ANS d. All of the above

2. On which data type, we cannot perform cluster analysis?

ANS d. None

3. Netflix's movie recommendation system uses

ANS c. Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is

ANS b. The tree representing how close the data points are to each other

5. Which of the step is not required for K-means clustering?

ANS d. None

6. Which is the following is wrong?

ANS c. k-nearest neighbor is same as k-means

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

i. Single-link

ii. Complete-link

iii. Average-link

ANS Options: d. 1, 2 and 3

8. Which of the following are true?

i. Clustering analysis is negatively affected by multicollinearity of features

ii. Clustering analysis is negatively affected by heteroscedasticity

ANS Options: a. 1 only

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?

ANS a. 2

10. For which of the following tasks might clustering be a suitable approach?

ANS b. Given a database of information about your users, automatically group them into different market segments.

11. Given, six points with the following attributes: Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

ANS a.

12. Given, six points with the following attributes: Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.

ANS b.

13. What is the importance of clustering?

ANS **IMPORTANCE OF CLUSTERING**

- Increased resource availability: If one Intelligence Server in a cluster fails, the other Intelligence Servers in the cluster can pick up the workload. This prevents the loss of valuable time and information if a server fails.
- Strategic resource usage: You can distribute projects across nodes in whatever configuration you prefer. This reduces overhead because not all machines need to be running all projects, and allows you to use your resources flexibly.
- Increased performance: Multiple machines provide greater processing power.
- Greater scalability: As your user base grows and report complexity increases, your resources can grow.
- Simplified management: Clustering simplifies the management of large or rapidly growing systems.