STATISTICS WORKSHEET-3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?
   a) Total Variation = Residual Variation – Regression Variation
   b) Total Variation = Residual Variation + Regression Variation
   c) Total Variation = Residual Variation * Regression Variation
   d) All of the mentioned

2. Collection of exchangeable binary outcomes for the same covariate data are called_____outcomes.
   a) random
   b) direct
   c) binomial
   d) none of the mentioned

3. How many outcomes are possible with Bernoulli trial?
   a) 2
   b) 3
   c) 4
   d) None of the mentioned

4. If Ho is true and we reject it is called
   a) Type-I error
   b) Type-II error
   c) Standard error
   d) Sampling error

5. Level of significance is also called:
   a) Power of the test
   b) Size of the test
   c) Level of confidence
   d) Confidence coefficient

6. The chance of rejecting a true hypothesis decreases when sample size is:
   a) Decrease
   b) Increase
   c) Both of them
   d) None

7. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

8. What is the purpose of multiple testing in statistical inference?
   a) Minimize errors
   b) Minimize false positives
   c) Minimize false negatives
   d) All of the mentioned

9. Normalized data are centred at____and have units equal to standard deviations of the original data

a) 0
b) 5
c) 1
d) 10

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What is Bayes' Theorem?

Ans.: The Bayes Theorem provides a principled way for calculating a conditional probability, or the likelihood of one event occurring if another has previously occurred. A conditional probability can lead to more accurate outcomes by including extra conditions — in other words, more data. In order to obtain correct estimations and probabilities in Machine Learning, conditional probabilities are required.

Although it is a powerful tool in the field of probability, Bayes Theorem is also widely used in the field of machine learning. Including its use in a probability framework for fitting a model to a training dataset, referred to as maximum a posteriori or MAP for short, and in developing models for classification predictive modeling problems such as the Bayes Optimal Classifier and Naive Bayes.

11. What is z-score?

Ans.: Z score is an important concept in statistics. Z score is also called standard score. Z-score is a statistical measure that tells you how far is a data point from the rest of the dataset. In a more technical term, Z-score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean. If a z-score is 0, it indicates that the data points score is identical to the mean score.

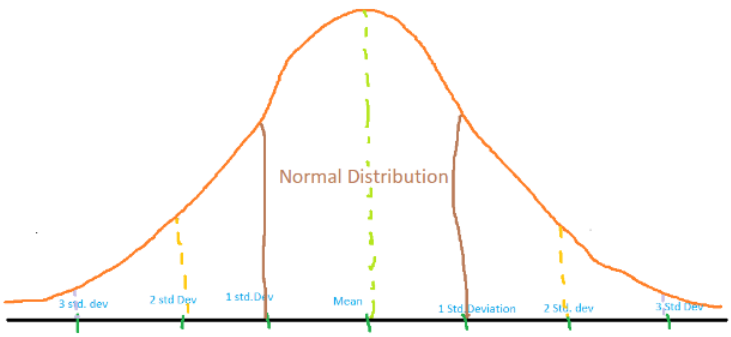
For example, a Z score of 2.5 means that the data point is 2.5 standard deviation far from the mean. And since it is far from the center, it's flagged as an outlier/anomaly.

Z-score is a parametric measure and it takes two parameters — mean and standard deviation. Once you calculate these two parameters, finding the Z-score of a data point is easy.

Z score = (x -mean) / std. deviation

Note that mean and standard deviation are calculated for the whole dataset, whereas x represents every single data point. That means, every data point will have its own z-score, whereas mean/standard deviation remains the same everywhere.

A normal distribution is shown below and it is estimated that
68% of the data points lie between +/- 1 standard deviation.
95% of the data points lie between +/- 2 standard deviation
99.7% of the data points lie between +/- 3 standard deviation



12. What is t-test?

Ans.: T Test is one of the foundational statistical tests. It is used to compare the means of two groups and determine if the difference is statistically significant. It is a very common test often used in data and statistical analysis. Also, t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.

$$t = \frac{\text{variance between groups}}{\text{variance within groups}}$$

If t-value is large => the two groups belong to different groups.
If t-value is small => the two groups belong to same group.

It is used in hypothesis testing, with a null hypothesis that the difference in group means is zero and an alternate hypothesis that the difference in group means is different from zero.

13. What is percentile?

Ans.: In statistics, a percentile is a term that describes how a score compares to other scores from the same set. While there is no universal definition of percentile, it is commonly expressed as the percentage of values in a set of data scores that fall below a given value. Percentiles are used in statistics to give you a number that describes the value that a given percent of the values are lower than.

14. What is ANOVA?

Ans.: ANOVA is a parametric statistical technique that helps in finding out if there is a significant difference between the mean of three or more groups. It checks the impact of various factors by comparing groups (samples) on the basis of their respective mean. It is 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.

H1: At least one mean of the groups are different.

15. How can ANOVA help?

Ans.: ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. 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.

ANOVA tests are a valuable tool in statistical analysis. They compare the difference between the means of two or more groups of data, and measure the degree to which levels or groups of an independent variable differ from each other. The purpose of an ANOVA test is to determine if there is a significant difference between the groups being studied by using variance. This is done using the 'f-test' in statistics to determine if the variables are significant. The ANOVA test allows for different levels of an independent variable to be compared against a dependent variable, which is beneficial in many different types of research.