STATISTICS WORKSHEET-3

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

Answer : Total Variation = Residual Variation + Regression Variation

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

a) random

b) direct

c) binomial

d) none of the mentioned

Answer: binomial

3. How many outcomes are possible with Bernoulli trial?

a) 2

b) 3

c) 4

d) None of the mentioned

Answer: 2

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

Answer: Type-I 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

Answer: Level of confidence

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

a) Decrease

b) Increase

c) Both of them

d) None

Answer: Decrease

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer: Hypothesis

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

Answer: 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

Answer: 0

10. What Is Bayes' Theorem?

Answer : Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself.

Bayes theorem is also known as the formula for the probability of “causes”. For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different colour balls viz. red, blue, black.

Bayes’ theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the

formula for the probability of “causes”. For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different colour balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability. In this article, let us discuss the statement and proof for Bayes theorem, its derivation, formula, and many solved examples.

11. What is z-score?

Answer: A z-score can be placed on a normal distribution curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). In order to use a z-score, you need to know the mean μ and also the population standard deviation σ.

Z-scores are a way to compare results to a “normal” population. Results from tests or surveys have thousands of possible results and units; those results can often seem meaningless. For example, knowing that someone’s weight is 150 pounds might be good information, but if you want to compare it to the “average” person’s weight, looking at a vast table of data can be overwhelming (especially if some weights are recorded in kilograms). A z-score can tell you where that person’s weight is compared to the average population’s mean weight.

The basic z score formula for a sample is:

$z = (x – \mu) / \sigma$

For example, let’s say you have a test score of 190. The test has a mean (μ) of 150 and a standard deviation (σ) of 25. Assuming a normal distribution, your z score would be:

$z = (x – \mu) / \sigma$

$= (190 – 150) / 25 = 1.6.$

The z score tells you how many standard deviations from the mean your score is. In this example, your score is 1.6 standard deviations above the mean.

12. What is t-test?

Answer : The t test is usually used when data sets follow a normal distribution but you don't know the population variance. For example, you might flip a coin 1,000 times and find the number of heads follows a normal distribution for all trials.

A t-test is a statistical calculation that measures the difference in means between two sample groups. The results from a t-test evaluate the significance of the mean difference to determine whether the outcomes occur by chance.

A t test is appropriate to use when you've collected a small, random sample from some statistical "population" and want to compare the mean from your sample to another value. The value for comparison could be a fixed value (e.g., 10) or the mean of a second sample.

t-tests are called t-tests because the test results are all based on t-values. T-values are an example of what statisticians call test statistics. A test statistic is a standardized value that is calculated from sample data during a hypothesis test.

There are three t-tests to compare means:

1. **a one-sample t-test**
2. **a two-sample t-test, and**
3. **a paired t-test**.

t-tests need three important data values: the standard deviation from each population group, the amount of data values from each group, and the mean difference between the values of the data sets.

The t test tells you how significant the differences between group means are. It lets you know if those differences in means could have happened by chance. The t test is usually used when data sets follow a normal distribution but you don't know the population variance.

13. What is percentile?

Answer: A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

The term percentile and the related term percentile rank are often used in the reporting of scores from norm-referenced tests. For example, if a score is at the 86th percentile, where 86 is the percentile rank, it is equal to the value below which 86% of the observations may be found. In contrast, if it is in the 86th percentile, the score is at or below the value of which 86% of the observations may be found. Every score is in the 100th percentile.
The 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3). In general, percentiles and quartiles are specific types of quantiles.

The range of values containing the central half of the observations is called the interquartile range: that is, the range between the 25th and 75th percentiles (the range including the values that are up to 25% higher or down to 25% lower than the median).

It is used with the median value to report data that are markedly non-normally distributed.

The percentile formula determines the performance of a person over others. The percentile formula is used in finding where a student stands in the test compared to other candidates. A percentile is a number where a certain percentage of scores fall below the given number.

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.

14. What is ANOVA?

Answer: 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.

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

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

The assumptions of the ANOVA test are the same as the general assumptions for any parametric test:

1. An ANOVA can only be conducted if there is **no relationship between the subjects** in each sample. This means that subjects in the first group cannot also be in the second group (e.g. independent samples/between-groups).
2. The different groups/levels must have **equal sample sizes**.
3. An ANOVA can only be conducted if the dependent variable is **normally distributed**, so that the middle scores are most frequent and extreme scores are least frequent.
4. Population variances must be equal (i.e. homoscedastic). **Homogeneity of variance** means that the deviation of scores (measured by the range or standard deviation for example) is similar between population

15. How can ANOVA help?

Answer: 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.

It provides the overall test of equality of group means. It can control the overall type I error rate (i.e. false positive finding) It is a parametric test so it is more powerful, if normality assumptions hold true.

Analysis of variance (ANOVA) is a tool to compare the means of several populations, based on random, independent samples from each population. It provides a statistical test to determine if population means are equal or not (i.e. came from the same distribution).

Real-world application of ANOVA test:

The researchers can take note of the sugar levels before and after medication for each medicine and then to understand whether there is a statistically significant difference in the mean results from the medications, they can use one-way ANOVA.

ANOVA's main advantage over t-tests is in **comparing multiple predictor variables at the same time**. This can make it easier to use, and faster.