STATISTICS WORKSHEET-3

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

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. b) Size of the test

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 centred at and have units equal to standard deviations of the original data
    Ans. a) 0

10. What Is Bayes' Theorem?
    Ans. The Bayes' theorem is a mathematical formula used to determine the conditional probability of events. In other words, the Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to that event.
    The Bayes' theorem is expressed in the following formula:

    $$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

    Where:

    - P(A|B) – the probability of event A occurring, given event B has occurred
    - P(B|A) – the probability of event B occurring, given event A has occurred
    - P(A) – the probability of event A
    - P(B) – the probability of event B

11. What is Z-score?
    Ans. Z-score also called a standard score is a numerical measurement that describes a value's relationship to the mean of a group of values. The Z-score is the number of standard deviations a given data point lies from the mean. It gives an idea of how far from the mean a data point is. If a Z-score is 0, it indicates that the data

point's score is identical to the mean score. For data points that are below the mean, the Z-score is negative & for the data points that are above mean, the Z-score is positive.

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).

The basic z score formula for a sample is:

$$Z = \frac{x - u}{\sigma}$$

Where:

Z = Z-Score
x = observed value
u = mean of the sample
σ = standard deviation of the sample

12. What is t-test?
    Ans. 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. A t-test compares the average values of two data sets and determines if they came from the same population or whether two groups are different from one another. It is also often used in hypothesis testing to determine whether a process or treatment actually influences the population of interest.

    T-test assumes a null hypothesis that the two means are equal.
    - The null hypothesis (H0) is that the true difference between these group means is zero.
    - The alternate hypothesis (Ha) is that the true difference is different from zero.

    Calculating T-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values.

    A large t-score, or t-value, indicates that the groups are different while a small t-score indicates that the groups are similar.

    The t-test assumes your data:

    - are independent
    - are (approximately) normally distributed.
    - have a similar amount of variance within each group being compared (a.k.a. homogeneity of variance)

    Types of T-tests: -

    i.   Paired t-test - If the groups come from a single population (e.g. measuring before and after an experimental treatment), perform a paired t-test.
    ii. Two-sample t-test - If the groups come from two different populations (e.g. two different species, or people from two separate cities) we will perform a two-sample t-test (a.k.a. independent t-test).
    iii. One-sample t-test - If there is one group being compared against a standard value (e.g. comparing the acidity of a liquid to a neutral pH of 7), perform a one-sample t-test.
    iv. Two-tailed t-test - If you only care whether the two populations are different from one another, perform a two-tailed t-test.
    v.   One-tailed t-test - If you want to know whether one population mean is greater than or less than the other, perform a one-tailed t-test.

13. What is percentile?

Ans. A percentile is a comparison score between a particular score and the scores of the rest of a group. In statistics, percentiles are used to understand and interpret data. The nth percentile of a set of data is the value at which n percent of the data is below it.

The percentile rank is calculated using the formula:

$$R = \frac{P}{100} * N$$

Where:
- P is the desired percentile
- N is the number of data (sorted from smallest to largest) points
- R is percentile rank

    i.   The 25th percentile is also called the first quartile.
    ii.   The 50th percentile is generally the median.
    iii.   The 75th percentile is also called the third quartile.
    iv.   The difference between the third and first quartiles is the interquartile range.

14. What is ANOVA?

Ans. ANOVA test (analysis of variance) is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance.

ANOVA splits the independent variable into 2 or more groups. For example, one or more groups might be expected to influence the dependent variable while the other group is used as a control group and is not expected to influence the dependent variable.

Types of ANOVA Tests:
- A one-way ANOVA has one categorical independent variable (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable. The one-way ANOVA test for differences in the means of the dependent variable.
- A two-way ANOVA has two or more categorical independent variables (also known as a factor), and a normally distributed continuous (i.e., interval or ratio level) dependent variable. A two-way ANOVA is also called a factorial ANOVA.

The formula for ANOVA is F = variance caused by treatment/variance due to random chance.

P-value is the probability of obtaining an F-ratio as large or larger than the one observed, assuming that the null hypothesis of no difference amongst group means is true.

- A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis,
- A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis.

The t-test determines whether two populations are statistically different from each other, whereas ANOVA tests are used when an individual wants to test more than two levels within an independent variable.

15. How can ANOVA help?

Ans. The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variables (such as the first example: age, sex, income). When you understand

how each independent variable's mean is different from the others, you can begin to understand which of them has a connection to your dependent variable and begin to learn what is driving that behaviour. ANOVA helps us to check the impact of one or more factors by comparing the means of different samples.