Statistics Worksheet 1 Solutions

1. Option a -> True
2. Option a-> Central Limit Theorem
3. Option b -> Modelling Bounded Count Data
4. Option d -> All of the mentioned
5. Option c -> Poisson
6. Option b -> False
7. Option b -> Hypohtesis
8. Option a -> 0
9. Option c -> Outliers Cannnot conform to the regression relationship
10. Normal distribution is a distribution in which the mean is zero and the standard deviation is 1. It forms a bell shaped curve. Normal distributions are symettric.For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two strandard deviatrions of the mean, and 99.7% are within three strandard deviations of the mean. Real life data rarely follows a perfect normal distribution. Normal Distribution has zero skewness. The normal distirbution has a kurtosis of three which indicates it has neither fat tails nor thin tails.
11. Ways of handling missing data->
    a. Deleting the entire row in which there is a cell with missing data
    b. Imputing The value->
        i. Replace all the missing data with an arbitary value
        ii. Replace missing data with the mean of the column
        iii. Replace missing data with the mode of the column
        iv. Replace missing data with the median of the column

v. Replace missing data with the previous value in the column(Forward Fill)
vi. Replace missing data with the next value in the column(backward fill)
vii. Interplolation

12. A/B testing, also known as split testing, refers to a randomized process wherein two or more versions of a variable are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and derives the business metric.

13. Mean imputaion is typically considerd a terrible practice since it ignores feature correlation. Conside the following scenario: we have a table with age and income, now suppose 21 year old has a missing value of the income, if we avergae the income of all people aged between 20 and 60, the 20 year old a higher income than what his actual income is.

    Mean imputation also decrease the variance of our data while increasing our bias.

14. Linear regression is used to predict the valuye of a variable on the value of another variable. The vaariable we want to predict is called the dependent variable and the variable we use to predict is called the independent variable. The goal of linear regression is to predict the slope and the Y-intercept of a line that predicts the relation between predictor variable and prediction variable.

    $Y = B0 + B1*X$

    Where B0 is the Y-intercept of the line and B1 is the slope

    The above is equation is the one we get from linear regression. The actual relation also has an error term which is independent of X and thus linear regression is unable to predict that error term

Actual Relation =

$y = b0 + b1x + \text{error term}$

The error term is generally irreducible error.

15. There are two branches of statistics ->

a. Descriptive Statistics -> It deals with the presentation and collection of data. This is usually the first part of the statistical analysis. In descriptive statistics we simply state what the data tells us.

b. Inferential Statistics - > It invloves drwaing the right conclusion from the statistical analysis that has been perfomed using descriptive statistics. It attempts to apply the conclusions obtained from the experiments to more general populations. This tries to answer quesitions about populations and samples that have not been tested in the given experiment.