Statistics Worksheet 1

1. a) True

2. d) all of the mentioned

3. c) Modelling contingency tables

4.

5. d) All of the mentioned

6. b) False

7. b) Hypothesis

8. a) 0

9. c) Outliers cannot conform to the regression relationship

10. It is also referred as the Gaussian distribution. It is symmetric about the mean, showing data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve at times that is why a normal distribution is the proper term for a probability bell curve. In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3. Normal distributions are symmetrical, but not all symmetrical distributions are normal. In reality, most pricing distributions are not perfectly normal.

11. According to data scientists, there are three types of missing data. Missing Completely at Random (MCAR) – that is, when data is completely missing at random across the dataset with no discernable pattern. Then there is Missing At Random (MAR) – that is when data is not missing randomly, but only within sub-samples of data. Finally, there is Not Missing at Random (NMAR), that is when there is a noticeable trend in the way the data is missing. I suggest two data imputation techniques that is, Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value

12. It is also referred to as split testing, which provides for a randomized experimentation process wherein two or more versions of variable, for example, web page, or page element, etc., are shown to different segments of website visitors at the same time to determine which version would leave the maximum impact and that would end up driving the business metrics.

13. The process of replacing null values in a data collection with the data's mean is known as mean imputation. It is typically considered terrible practice since it ignores feature correlation. It decreases the variance of our data while increasing bias. Going deeper into mathematics, a smaller variance leads to the narrower confidence interval in the probability distribution and therefore, it leads to nothing else than introducing a bias to the model.

14. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. It fits a straight line that minimizes the discrepancies between predicted and actual output values. It estimates the value of X (dependent variable) from Y (independent variable).

15. Data collection, Descriptive statistics and Inferential statistics.