Assignment 1 – Statistics Worksheet

1) True
2) Central Limit Theorem
3) Modeling bounded count data
4) All of the mentioned
5) Poisson
6) False
7) Hypothesis
8) 0
9) Outliers cannot conform to the regression relationship

10) Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that 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.

    - 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) When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.

    - The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

- The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

12) An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

    To put this in more practical terms, a prediction is made that Page Variation #B will perform better than Page Variation #A. Then, data sets from both pages are observed and compared to determine if Page Variation #B is a statistically significant improvement over Page Variation #A.

13) Is mean imputation of missing data acceptable practice?

    Ans) No it is not acceptable as Mean imputation does not preserve the relationships among variables. Mean Imputation Leads to An Underestimate of Standard Errors.

14) Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

    (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

    (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

    These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

    The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated

dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15) The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

- Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

- Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.