Statistics – WORKSHEET

1. a) True
2. a) Central Limit Theorem
3. b) Modeling bounded count data
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) 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. In this the mean =median=mode and the standard deviation is 1.

11. To handle missing values in dataset we use mean & mode when the data is small. The mean is used in continuous types of features while mode is used in discrete or ordinal type of feature variable. If the dataset is big & contains less number of null values it's better to remove such records.
    The most important missing data imputation techniques for handling missing data during prediction time are reduced feature models, distribution-based imputation, prediction value imputation.

12. A/B testing is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

13. Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should.

    Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

14. In statistics, linear regression is a linear approach for modelling the relationship between a dependent variable (target value) & independent variable (Features in data). When the independent variable is one then the process is known as Simple Linear Regression. And for more than one independent variable the process is known as multiple Linear Regression.

15. The branches of Statistics are:

    Descriptive Statistics: In this data can be described without any statistical tool like marks in class, Height of student.
    Inferential Statistics: In this data is too big to find terms like we take average of samples from population like for prediction of the vote election.