STATISTICS– WORKSHEET 1

1. (a) True
2. (a)Central limit theorem
3. (c) Modeling bounded count data
4. (d)All 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.
11. There are several ways to handle missing data:

a. Delete rows with missing data
b. Mean/Median/Mode imputation
c. Assigning a unique value
d. Predicting the missing values
e. Using an algorithm which supports missing values, like random forests.

The easiest and quickest approach to a missing data problem is dropping the offending entries. This is an acceptable solution if we are confident that the missing data in the dataset is missing at random, and if the number of data points we have access to is sufficiently high that dropping some of them will not cause us to lose generalizability in the models we build (to determine whether or not this is case, use a learning curve).

12. A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal. Running an A/B test that directly compares a variation against a current experience lets you ask focused questions about changes to your website or app and then collect data about the impact of that change.
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-yearold will appear to have a much higher fitness score that he actually should.
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.
15. The two types of statistics have some important differences.

a. Descriptive statistics is the type of statistics that probably springs to most people's minds when they hear the word "statistics." In this branch of statistics, the goal is to describe. Numerical measures are used to tell about features of a set of data.
b. Inferential statistics are produced through complex mathematical calculations that allow scientists to infer trends about a larger population based on a study of a sample taken from it