STATISTICS WORKSHEET-1

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 called **Gaussian distribution**, the most common distribution function for independent, randomly generated variables. Its familiar bell-shaped curve is ubiquitous in statistical reports, from survey analysis and quality control to resource allocation. The graph of the normal distribution is characterized by two parameters:

    the **mean**, or average, which is the maximum of the graph and about which the graph is always symmetric; and the **standard** deviation, which determines the amount of dispersion away from the mean.
11. **Handle missing data** :

    1. **Listwise Deletion**: Delete all data from any participant with missing values. If your sample is large enough, then you likely can drop data without substantial loss of statistical power. Be sure that the values are missing at random and that you are not inadvertently removing a class of participants.

    2. **Recover the Values**: You can sometimes contact the participants and ask them to fill out the missing values. For in-person studies, we've found having an additional check for missing values before the participant leaves helps.

    3. **Educated Guessing**: It sounds arbitrary and isn't your preferred course of action, but you can often infer a missing value.

    4. **Average Imputation**: Use the average value of the responses from the other participants to fill in the missing value.

    5. **Common-Point Imputation**: For a rating scale, using the middle point or most commonly chosen value.

    6. **Regression Substitution**: You can use multiple-regression analysis to estimate a missing value. We use this technique to deal with missing SUS scores. Regression substitution predicts the missing value from the other values. In the case of missing SUS data, we had enough data to create stable regression equations and predict the missing values automatically in the calculator.

    7. **Multiple Imputations**: The most sophisticated and, currently, most popular approach is to take the regression idea further and take advantage of correlations between responses. In multiple imputation [pdf], software creates plausible values based on the

correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions.

12. **A/B testing** is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment
13. Mean imputation is typically considered terrible practice since it ignores feature correlation and decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.
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. There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.
Data collection is all about how the actual data is collected.
Descriptive statistics is the part of statistics that deals with presenting the data we have.
Inferential statistics is the aspect that deals with making conclusions about the data.