Statistics Worksheet

1. True
   Bernoulli distribution is a distribution which has only two outcomes: success or failure. Success is taken as 1 and failure is takes as 0.

2. Central Limit Theorem
   The central limit theorem becomes normal distribution when the data is sufficiently large.

3. Modeling bounded count data
   The Poisson distribution is applied on the data where the average time between the events is known but the exact time of the event is not known. For example, a sudden movement in a particular stock price.

4. All the statements are correct.

5. Poisson
   Poisson random variables are used to count the occurring of an event in a given interval of time.

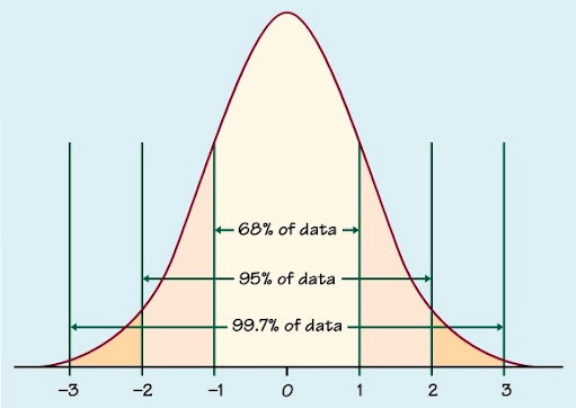
6. False, replacing the standard error with its estimated value doesn't change the central limit theorem.

7. Hypothesis
   Hypothesis testing is used to make decisions on the basis of data. For example if we are testing the Drug A and Drug B on a group of people, we can't reject the null hypothesis if small random changes can change the result such doing more exercise, or taking a healthy diet.

8. 0
   The normalized data are centered at the 0. The normalization is used in data science field when we have features of different scales or we have to compare data of two different scales. For example, the exam held in shift 1 is relatively easier than the exam held in shift 2. Two compare the marks distribution of these two shifts, we have to normalize the data.

9. Outlier can conform the regression relationship.
   Outliers are the data points which are significantly different from the remaining data. These outliers can make the regression highly biased towards.

10. It is a probability distribution which shows that the data near the mean are more frequent in occurrence than the data far from the mean. This distribution is symmetrical about the mean. The normal distribution is also known as Gaussian distribution which looks like a bell curve as shown below.



The standard distribution has two important parameters: mean(μ) and standard deviation(σ). It is also shown in the graph that the 68% of the data lies within $\pm\sigma$, 95% of the data lies within $\pm 2\sigma$ and 99.7% of the data lies between $\pm 3\sigma$ from the mean. As the data distribution is symmetrical, the skewness of the data is zero and its kurtosis is 3.

11. The quickest way to handle the missing values is to delete those rows but by doing this we can lose a lot of information. If a particular column has a significant number of missing values that column can be dropped. The other method to handle the missing values is to impute the missing values with mean or median or mode. If the missing values are categorical, then it can be replaced by the most frequent category. If the missing values in a categorical variable are significantly high then we can fill the missing values with a new category such as "unknown". These methods just save us from losing the information. We have better methods to impute the missing values. We can use the algorithm based approach such as k-NN. In this method, the missing values are imputed based on the other feature available. Similarly, machine learning based approach can also be applied such as linear regression, Random Forest etc. Every real life dataset has missing values. One shouldn't restrict to fill the missing values with one method. Different methods can be applied to get the robust model. Having domain knowledge of the data would be helpful to choose the correct method of imputation. There are several libraries available in the python such as fancyimpute, impyute etc.

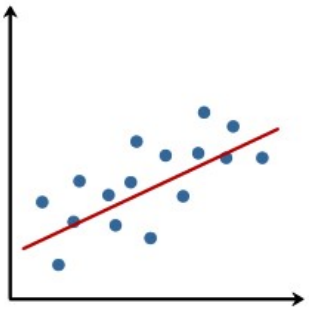
12. A/B testing is the act of running a simultaneous experiment between two or more products to see which one performs the better than the others. The testing is not limited to the products; it can be used for different images of a product to get CTR, different page layout to attract more traffic. A/B test use to determine the impact of certain change that is relatively inexpensive. A/B test can be conducted on several verticals such as email campaign, web page, running advertisements etc.

13. Imputing with mean not suitable for all dataset. Imputing with mean ignores the correlation between the other variables. The other problem with the mean imputation is that when dataset is small, it will change the variance significantly. This significant change in the data leads to a biased model. Alternative approach such as k-NN, Random Forest can reduce this problem and model will perform better.

14. Linear regression set a linear relation between the one or more independent variables with one dependent variable. This is basically done by adjusting the coefficients of different variables which are also know as weights. Mathematically it can be represented as

$$Y = AX$$

Where Y is dependent variable, A is weight matrix and X is variable matrix. A is a row matrix and X is a column matrix. The most common method to best fit a line with the data is least square



method. This method fit the best line by minimizing the square of deviation of each data point. The coefficient matrix also gives us the importance of a particular variable. Higher the weight of the particular variable, more important the variable is.

Linear regression has these five assumptions:

1. Linear relationship
2. Multivariate normality
3. No or little multicollinearity
4. No auto-correlation
5. Homoscedasticity

15. There are two branches of statistics: descriptive statistics and inferential statistics.

Descriptive statistics: This branch of statistics focused on collecting, summarizing, and presenting the data. In this, we determine the mean, median, mode, standard deviation, central tendency etc. Graphs can be used to visualize the data. For example, analyzing the marks distribution of a class by determining average marks, maximum marks etc.

Inferential statistics: The branch of statistics in which we take a sample of the data and draw a conclusion for the population data. This technique is used by the statistician for data analyzing, making conclusions from the limited information.

Different types of inferential statistics:

1. Regression analysis
2. Analysis of variance (ANOVA)
3. Analysis of covariance (ANCOVA)
4. Statistical significance (t-test)
5. Correlation analysis