STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

b) Modeling bounded count data

4. Point out the correct statement.

d) All of the mentioned

5. _____ random variables are used to model rates.

c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

b) False

7. 1. Which of the following testing is concerned with making decisions using data?

b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0

9. Which of the following statement is incorrect with respect to outliers?

c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?

→ Normal /Gaussian Distribution is a kind of distribution where in if a random variable X follows a gaussian distribution and we try to plot it with a histogram, then the distribution would be in the form of a Bell Curve. The centre point of the Bell Curve would be the median (**μ**).If we go one position to right then it would be **μ+σ**, if we go 2 position towards right then it would be **μ+2σ** and if we go 2 position towards right then it would be **μ+3σ**. Similarly, if we go towards the left from **μ**, the corresponding values would be **μ-1σ** for 1 position, **μ-2σ** for 2 position and **μ-3σ** for 3 positions.

The values between **μ+σ** and **μ-σ** would be the range of 1st standard deviation, the values between **μ+2σ** and **μ-2σ** would be the range of 2nd standard deviation and the values between **μ+3σ** and **μ-3σ** would be the range of 3rd standard deviation.

If the random variable **X** follows a normal distribution, then,

1. Approximately 68% of data points of **X** would fall under the range of 1 standard deviation (**μ-σ <= x <= μ+σ)**.
2. Approximately 95% of data points of **X** would fall under the range of 2 standard deviation (**μ-2σ <= x <= μ+2σ)**.
3. Approximately 99.7% of data points of **X** would fall under the range of 3 standard deviation (**μ-3σ <= x <= μ+3σ).**

11. How do you handle missing data? What imputation techniques do you recommend?

→ We could basically handle the missing values in 3 ways:

1. **Deleting the records that have missing values**: If we have a huge data set which has a with a very few missing values, we could delete the record having Null values.
2. **Create a separate model to handle the missing values:** Creating a Separate Model to handle missing values takes a lot of time. In this method, we divide the dataset into 2 parts the one with non-missing values in a particular feature to be a training dataset and the part of dataset that is missing the values in a particular feature to be a test data set and use the predictions to fill the missing values. In this we need to create separate models for missing values in each feature and need to dataset to be big for the model to learn properly.
3. **Use Statistical methods (Mean/ Median / Mode) :** Statistical methods is probably the best method.

The imputation technique that I would recommend would be Statistical methods such as Mean/Median or Mode if the highest degree of accuracy of the data is not a concern and if it is then I'd create a ML model to substitute the missing values.

12. What is A/B testing?

→ A/B Testing is nothing but creating 2 models/ad/thumbnail…etc for the same Dataset/Product and checking which model/ad is performing better off the two.

Eg. A product based company is launching a new product they may make 2 ad's for the product, run a beta version of the ad and see which ad off the 2 are the people more inclined towards this may be by view time, clicks etc so that in the production they could analyse and release the ad which consumers prefer and thus having better chance of selling the product.

13. Is mean imputation of missing data acceptable practice?

→ Yes, until and unless the missing data is not a categorical data, this may be one of the fastest ways to handle the data but its not very accurate as it does not take into account any co-relations or account for any uncertainty in the imputations.

14. What is linear regression in statistics?

→ It is a statistical way of measuring the relationship between variables.

The Equation of linear equation is:

$$y=mx+c$$

where,

y = the value to be predicted

m = the slope or constant

x = Input

c = Intercept/Bias

15. What are the various branches of statistics?

→ There are **two main branches** of statistics
- Inferential Statistic.
- Descriptive Statistic.