Statistics – Worksheet 1

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

Ans: 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?

Ans: A. Central Limit Theorem

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

Ans: B. Modeling Bounded Count Data

4. Point out the correct statement.

Ans: D. All of the mentioned

5. ______ random variables are used to model rates.

Ans: C. Poisson

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

Ans: B. False

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

Ans: B. Hypothesis

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

Ans: A. 0

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

Ans: C. Outliers cannot conform to the regression relationship

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

Ans: Normal distribution also known as Gaussian distribution. Its widely used in Data science and machine learning project with unexpected real-world scenarios. Based on Gauss's work, Pierre Simon LaPlace, introduced the Central Limit Theorem (CLT). CLT is an essential concept of statistics which represent the behaviour of independent, random variable inclined towards the normal distribution. Normal distribution has several characteristics. The two most important parameters are Mean and Standard Deviation. In normal distribution the mean and stdev are not fixed. They can take on any value. Normal distribution gives you a symmetrical bell curve where maximum data is centred around one point and some data will fall away towards the two opposite ends.

There is one very important rule for standard Normal deviation known as Empirical rule:68/95/99.7. This rule plays a significant role during Hypothesis testing. This rule states that 68% observations are within ±1 stdev from the mean, 95% observations are within ±2 stdev from the mean, and 99.7% observations are ±3 stdev from the mean. Values outside ±3 stdev are considered as Outliers. What it means is values which are further away from means are less likely to be part of the distribution itself. Normal distribution can be transformed into standard normal distribution anytime.

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

Ans: Missing Data can create incorrect analysis due to the missing of relevant information. Output or good results are often depending upon the quality of data. There are three main types of missing data.

I. First is MCAR (Missing completely at Random): this analysis assumes that the missing data is unrelated to any unobserved data. This provides unbiased & reliable estimates.
II. second is MAR (Missing at Random): this is more common than MCAR.
III. the third one is MNAR (Missing not at random).

we can try the above-mentioned data analysis strategies for missing data. Listwise deletion, Imputation, Multiple imputation, & Full information Maximum likelihood.

Multiple imputation and Maximum Likelihood are commonly used because they assume missingness is random and cannot be ignored in the modelling process.

12. What is A/B testing?

Ans: A/B Testing is a method to compare and analyse two versions of variables to find out which one performs better in a controlled environment. A/B testing is popular and widely used statistical tools. In A/B testing one need to form Two Hypothesis.

- Null Hypothesis(H0) and: this will assume no difference will appear in results if we change one variable.
- Alternative Hypothesis (Ha): this is exactly the reverse of null hypothesis stating changing one variable can impact outcome.

In these testing it is important to know whether the results are Statistically Significant or not. It means we have to analyse if the event has happened as a result of chance. To check the statistical significance, we use two parameters. P-value and Alpha value. We can use P-Value to justify the Null Hypothesis. It means the lower the p-value compared to the Alpha value, the more Unproven the null hypothesis becomes. Alpha value is set generally prior to the experiment to 0.05 or 0.01 depending on the experiment. In other words, P-Value & Alpha value are being compared with each other to identify which hypothesis is true. To know the final outcome, we have to give Confidence interval, it is an observed range where all the results of the test will be seen. Confidence level is directly opposite with the P-value. If p-value is set at 0.05%, then confidence level would be set manually at 95%.

13. Is mean imputation of missing data acceptable practice?

Ans: Imputation is a process in statistics where missing data is replaced by the substitute values. There are several ways to address missing data. In mean imputation method null values are getting replaced with the mean of overall data.

This procedure is not considered acceptable practice because it does not consider the correlations between features. Thus, it becomes problematic for multi-variable /multi-categorical analysis. As I said earlier there are several approaches to deal with missing data. All has pros and cons but mean imputation is far worse because it provides less accurate data and can be perceived as providing biased output.

14. What is linear regression in statistics?

Ans: Linear regression is an analysis technique to predict the value of a variable using another variable. The variable we want to predict is dependent variable and other variable which we are using for prediction is known as independent variable. Linear regression draws a best fit straight line that minimised the discrepancies between predicted and actual results. The main goal to perform linear regression is to identify the effect that the independent variable has on dependent variable. Linear regression is a good strategy to forecast change or effects and predict future trends. There are various types of linear regression. To name a few there is Multiple linear regression, Logistic regression, Multinomial regression etc.

15. What are the various branches of statistics?

Ans: statistics is the branch of applied mathematics that involves the collection, description, organization, interpretation & presentation of data. Thera re three real branches of statistics

I. Data collection: data collection is all about the data collected for analysis

II. Descriptive statistics: descriptive statistics is all about the presentation of the collected data. The main purpose od descriptive statistics is to present the data such a way that anyone can understand it.

III. Inferential statistics: inferential statistics deals with making conclusions about the available data. As the name inferential suggests it’s about making decision or drawing conclusion basis of evidence and reasoning.