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 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.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

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 centred 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?

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 graphical form, the normal distribution appears as a "bell curve". In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

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

Different ways of handling missing data

Zero Replacement: Here, we can replace the missing value with zero irrespective of everything.

Min or Max Replacement: Replace the missing value with the minimum or maximum value of a feature.

Mean/ Median/ Mode Replacement: Replace missing value with mean or median or most frequent feature value.

Also, one can replace the value of the missing cell with the previous cell's value. This kind of technique is popular while inputting time series data. For example, if the price of an instrument is missing on the n-th day, we can replace it with the (n-1)th day's price.

12.What is A/B testing?

A/B testing is an experiment on two variants to see which performs better based on a given metric. Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions.

A/B testing is also a form of statistical and two-sample hypothesis testing. **Statistical hypothesis testing** is a method in which a sample dataset is compared against the population data. **Two-sample hypothesis testing** is a method in determining whether the differences between the two samples are statistically significant or not.

13.Is mean imputation of missing data acceptable practice?

Yes

14.What is linear regression in statistics?

A **linear regression model** is used to depict a relationship between variables which are proportional to each other. Meaning, the dependent variable increases/decreases with the independent variable.

In the graphical representation, it has a straight linear line plotted between the variables. Even if the points are not exactly in a straight line (which is always the case) we can still see a pattern and make sense out of it.

Example: As the age of a person increases, the level of glucose in their body increases as well.

A **multiple regression model** is used when there is more than one independent variable affecting a dependent variable. While predicting the outcome variable, it is important to measure how each of the independent variables moves in their environment and how their changes will affect the output or target variable.

Example: Chances of a student failing their test can be dependent on various input variables like hard work, family issues, health issues, etc.

15.What are the various branches of statistics?

Statistics have majorly categorised into two types:

1 .Descriptive statistics

2. Inferential statistics

Descriptive Statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.