Sachin Boora (DS2302)

https://github.com/sachinboora/Internship

Worksheet_set_1

Statistics Worksheet

Ans 1 : True, Bernoulli random variable takes only 1 value either 0 or 1.

Ans 2 : Option A is correct - Central limit theorem

Ans 3 : Poisson Distribution is used to model time/event data such as number of phone calls received in an hour. Modelling bounded count data means there is a certain limit to data which is not appropriate for use of Poisson Distribution.

Ans 4 : Option C is correct - The square of standard normal random variable follows chi-squared distribution.

Ans 5 : Poisson distribution is used to model rates so option C is correct.

Ans 6 : True

Ans 7 : Hypothesis testing is concerned with making decisions using data.

Ans 8 : Option A - 0

Ans 9 : Outliers can conform to regression relationship so option C is incorrect.

Subjective Questions

Ques 10 : What do you understand by the term normal distribution?

Ans 10 : It is a probability distribution commonly used in statistical analysis. It is symmetric around its mean.

The Normal distribution is defined by two parameters:

A) Mean
B) Standard Deviation

Mean represents the centre of distribution while Standard Distribution measures the spread or variability of the distribution.

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

Ans 11 : Handling missing data is an important task in data analysis as it can affect the whole result of the analysis or can give answers which are not expected. There are many ways we can handle missing data which includes case analysis and different imputation techniques.

There are many imputation techniques like mean imputation, regression imputation, multiple imputation.

However, multiple imputation is what I recommend to use as it provide most accurate estimate and can account for uncertainty in the imputation process.

Ques 12 : What is A/B Testing?

Ans 12 : Also known as split testing, is a statistical method used to compare two versions of a product. It is a commonly used testing in marketing, UX Designs.

The two versions are assigned to random users and their behaviour is monitored to see which version perform better. The results of test are the analysed using various statistical methods.

Ques 13 : Is mean imputation of missing data a acceptable practise ?

Ans 13 : Mean imputation is a simple and commonly used method for handling missing data however it has some drawbacks and may not always be an acceptable practise.

One of the main drawback it has is that it assumes that missing value has same distribution as the non missing values in the same variable.

It also ignores the uncertainty associated with missing data.

Ques 14 : What is Linear Regression in statistics?

Ans 14 : It is a statistical method used to model the relationship between two continuous variable, where one variable is independent variable and other is depend variable. It is a commonly used technique in data analysis.

This model assumes that there is a linear relationship between two variables which can be represented by a straight line on a scatter plot.

The goal of linear regression is to find the best fitting line that represents the relationship between two variables.

Ques 15 : What are various branches of Statistics ?

Ans 15 : Statistics can be divided into various branches. Some of them are :

A) Descriptive Statistics
B) Inferential Statistics
C) Probability Theory
D) Bayesian Statistics
E) Multivariate Statistics
F) Non Parametric Statistics
G) Time Series analysis.