internship No : DSC2311

STATISTICS WORKSHEET-1

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

Ans :- `True`

Question 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 :- `Central Limit Theorem`

Question 3:- Which of the following is incorrect with respect to use of Poisson distribution

Ans :- `Modeling bounded count data`

Question 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) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

Ans :- d) `All of the mentioned`

Question 5:- random variables are used to model rates.

Ans :- `Poisson`

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

Ans :- `False`

Question 7:- Which of the following testing is concerned with making decisions using data

Ans :- `Hypothesis`

Question 8:- Normalized data are centered at and have units equal to standard deviations of the original data.

Ans :- 0

Question 9 :- Which of the following statement is incorrect with respect to outliers?

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Ans :- `c) Outliers cannot conform to the regression relationship`

Question 10:- What do you understand by the term Normal Distribution?

Ans:- Normal distribution is evenly distributed across its mean

Also known as Bell curve as well as Gaussian distribution, which have half of the data at right of the mean and half of the data at left of the mean.

Mean, Median and Mode are all equal to one another. Moreover, these values all represent the peak, or highest point, of the distribution. The distribution then falls symmetrically around the mean, the width of which is defined by the standard deviation.

About 68 of all cases within 1 STD of the mean, about 95% of cases within 2 STD of the mean

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

Ans :- There are multiple ways to dealing with missing data, the most widely used are

Mean or Median Imputation :- This process consists of replacing all occurrences of missing values (NA) within a variable by the mean or median.

Multivariate Imputation :- It assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. They use predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables.

Question 11 :- What is A/B testing?

Ans:- A/B testing is a methodology for testing product changes. The users are split into two groups - the control group which sees the default feature, and an experimental group that sees the new features

How A/B tests are conducted:

- ⇨ How subjects are randomized into treatments
  - ⇨ Who in the organization is running the A/B test
  - ⇨ Who is in charge of the analysis

4 challenges for designing a good A/B test:

1. Deciding what to test
2. Deciding which subjects to target
3. Modelling the data
4. Choosing the sample size

Question 13:- Is mean imputation of missing data acceptable practice?

Ans:- Yes it's an acceptable practice based on certain criteria, by using mean imputing, it keeps the mean of the observed data intact. So if the data are missing completely at random, the estimate of the mean remains unbiased. By imputing the mean, thus helps to keep the sample size up to the full sample size But the challenges associated with them are

- ⇨ Mean imputation reduces the variance of the imputed variables.
- ⇨ Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- ⇨ Mean imputation does not preserve relationships between variables such as correlations

Question 14 :- What is linear regression in statistics?

Ans:- Linear regression indicates the relationship between one or more features and labels. Linear regression is commonly used for predictive analysis and modelling, like sales vs. advertisement data etc.

- **Normally explained by the straight line equation**

Y = MX+C

Where Y = Label (target as sales) X = Feature (factor like mode of ad)

M = Coefficient /slope

C= Intercept

- **Types of Linear Regression**

1. **Simple linear regression**
   1 dependent variable, 1 independent variable

2. **Multiple linear regressions**
   1 dependent, 2+ independent variables

Question 15:- What are the various branches of statistics?

Ans:- The study of data via statistics can be widely divided into two categories

a) Descriptive statistics
b) b) Inferential statistics

Descriptive Statistics

The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

Inferential Statistics

The branch of statistics that analyzes sample data to draw conclusions about a population