STATISTICS WORKSHEET-1

Q1 to Q9 has only one correct answer. Choose the correct option to answer your question.

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

a) True
b) False

Answer. (a) True
As a result of a binary outcome, the Bernoulli distribution is generated.

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
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

Answer. (a) Central Limit Theorem

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

a) Modeling event/time data
b) Modeling bounded count data
c) Modeling contingency tables
d) All of the mentioned

Answer. (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) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned

Answer. (d) All of the mentioned.

5. _____ random variables are used to model rates.

a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Answer. © Poisson Distribution

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

a) True
b) False

Answer. (b) False

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

a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Answer. (b) Hypothesis

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

a) 0
b) 5
c) 1
d) 10

Answer. (a) 0

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

Answer. © Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer. Normal distribution holds a bell-shaped curve with mean, median, mode - all are equals to each other. There is no dispersion in the data and standard deviation value holds between + and - 1.

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

Answer. We can handle the missing data either with the PANDAS library or SimpleImputer technique of Scikit Learn library.

Let's suppose df is the dataset which is loaded in the Jupyter Notebook.

The syntax for the Pandas is:

- **Df = df.isnull().sum() #With this, we will know whether missing data exist or not. If yes, then further it can be removed with the drop.na() command or can be filled with the mean of relevant values.**
- **To remove the missing values, the command is: df.dropna(axis=0)**
- **To fill the missing values, the command is: df.fillna(). Here we can apply backward or forward techniques.**

The syntax for using the SimpleImputer Technique of Scikit Learn Library is:

- **From sklearn.impute import SimpleImputer**
- **Imp = SimpleImputer(strategy = "most_frequent")**
- **Df['column_1'] = imp.fit_transform(df['column_1'].value.reshape(-1,1))**
- **#Here NAN values will be changed with the mode of the data.**

12. What is A/B testing?

Answer. A/B testing refers to the split testing. In this method, two variants of a web page are displayed to different segments of website visitors simultaneously and their conversion rates are compared.

13. Is mean imputation of missing data acceptable practice?

Answer. No it is not acceptable practice because it ignores the very significant feature, that is, correlation. It is important to know how the variables are correlated to each other. But mean imputation ignores that.

14. What is linear regression in statistics?

Answer. Linear regression is about finding the best fit line. Best fit line means the difference between the actual and predicted output is minimum.

It is used when the data is given in decimal values like home price, doctor fee, and many more. Here my output is continuous in nature. X is an independent variable and Y is a dependent variable. In linear regression, coefficient shows how much input contribution there is in "y" output. Errors help in showing the difference between predicted and actual output.

15. What are the various branches of statistics?

Answer. In general, there are two statistics of branches: descriptive statistics, and inferential statistics.