WORKSHEET

STATISTICS WORKSHEET-1

Q1 to Q9 have 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)

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: c(Poission)

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

Answer: b(False)

7. 1. 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. 4. 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: c(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?

It is also called as Gaussion Distribution.
It is mostly commonly seen on continues distribution in nature.
e.g Age, Marks, Salary etc.
Just as a binomial distribution every event is independent from one another.
Mean, Median and Mode are lie up in a centre of the distribution of the mean.
The normal distribution is in bell shape or bell curve.

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

In real time scenario we get lots of different types of data. Some data may be blacked or missed. To hadle such type of data we have some imputational techniques and there are:

- Mean
- Median
- KNNImputer
- SimpleImputer
- IterativeImputer

If you want to impute a value in a categorical column with mean method it does not give a proper accurate value to that position.

e.g Suppose a column is categorized in binary 1 & 0 if we are trying to imputing with a mean value of that column. It gives a float value which is not currect. At this time we can use mode method. Mode is nothing but a high number of accurance in a respective column.

KnnImputer is used fill in a missing value. It is used to predict missing values to each missing points based on n_neighbours and make a mean of that all point which fall under in n neighbours.

12.What is A/B testing?

A/B testing one of the important step in a data science cycle. A/B testing is an example of statistical hypothesis testing . The process of hypothesis is about a relationship between to different data sets. Those dataset are then compared against on each other to determine whether there is statistically significant relationship or not.

Example :
Prediction has done on average marks of one classroom before appointing a new class teacher and average marks of classroom after appointing a new class teacher. Then after both the average marks are observed and compared if any statistical sifnificant over the average marks.

13.Is mean imputation of missing data acceptable practice?

It’s not a good practise to impute missing values by using mean method in every problems. Suppose out of 100 rows 25 rows are misssing data which is in countinues nature. In this situaltion we can use a mean method but it replace mean value to entire 20 rows which is not going to give real values to the each poins and it is not a proper solution. Imputing missing value are our imaginary value. While imputing a value it should be close to the real value. Rather using mean method there are soo many other techniques based on problem we can use such as KnnImputer, IterativeImputer etc. These are best techniques to use to impute missing value based on type of problems.

14.What is linear regression in statistics?

Linera Regression is a relationship between two variable. One variable is dependent on another variables which is independent. If independent are single variable then we can called as simple linear regression or more than one variables are independent then called as mutliple linear regression. It is an approch to see how affects on dependent variable by changing the independent variable. Linear regression is a basic and commonly used type of predictive analysis.
It is defined by formula : **$\underline{y = mx+ c}$**
where
y= dependent variable.
m=coefficient / slope
x=independent variable
c=intercept
It is a method to find a best linear realationship between dependent and independent variable.

15.What are the various branches of statistics?

Statistic is nothing but a some calculation done on a data to get some isight information and to use for normalise the data, clean the data. Statistic can be devided into two branches and there Descriptive and Inferential statistic.
If we able to describe the data of each individual of each data point then it comes under the part of descriptive statistics.
e.g Average marks of one classroom having 30 to 60 number of students.

If the data is too big hard to describe each data point from the data. Let's consider whole data as a population and took random data points from entire population and say it as sample data.
Work, analyze and make predictions on this sample data and inference to entire data i.e population.

e.g Average marks of multiple classrooms each classroom having 30 to 60 number of students.