**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

Ans. 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

Ans. 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

Ans. 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

Ans. All of the mentioned

5. \_\_\_\_\_\_ random variables are used to model rates.

a) Empirical b) Binomial

c) Poisson d) All of the mentioned

Ans. Poisson

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

a) True

b) False

Ans. 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

Ans. 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

Ans. 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

Ans. None of the mentioned (Outliers can 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?**

Ans. A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme. 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.

Height is one simple example of something that follows a normal distribution pattern. Most people are of average, the numbers of people that are taller and shorter than average are fairly equal and a very small number of people are either extremely tall or extremely short.

Normal distribution curves are somethings designed with a histogram inside the curve. The graphs are commonly used in mathematics, statistics and corporate data analytics.

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

Ans. Imputation techniques :

1. **Missing at Random** : Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.
2. **Missing Completely at Random** : The fact the a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
3. **Missing not at Random** : Two possible reasons are the missing value depends on the hypothetical value or missing value us dependent on some other variable’s value.

**12. What is A/B testing?**

Ans. A\B testing is basically use to compare to two different products. Here, we’ll understand what A\B testing is and how you can leverage A\B testing in data science using Python.

It is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. Here, either you can use random experiments, or you can apply scientific and statistical methods. It is a widely used concept in most industries nowadays, and data scientist are at the forefront of implementing it. In this article, I will explain A\B testing in-depth and how a data scientist can leverage it to suggest changes in a product.

It’s points to the combination of elements that helps keep visitors on site or app longer. The more time visitors spend on site, the likelier they’ll discover the value of the content, ultimately leading to a conversion. It includes application of statistical hypothesis testing or “two-sample hypothesis testing” as used in the field of statistics.

**13. Is mean imputation of missing data acceptable practice?**

**Ans.** True, imputing the mean preserves the mean of the observed data. So, if the data are missing completely at random, the estimate of the mean remains unbiased. The process of replacing null values in a data collection with the data’s means in known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation.

**14. What is linear regression in statistics?**

**Ans.** In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables are also known as dependent and independent variables. Simple linear regression is a regression model that estimates the relationship between one dependent variable using a straight line. Both variables should be quantitative.

Linear regression most often uses mean-square error to calculate the error of the model. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula:- y = c + b \* x

Where y = estimated dependent variable score

c= constant,

b= regression coefficient

x= score on independent variable



**15. What are the various branches of statistics?**

**Ans.** There are three branches of statistics are:- a) Data Collection

b) Descriptive Statistics

c) Inferential Statistics

a) Data Collection : Data collection is all about how the actual data is collected. In such data marks in a class test, this is fairly straightforward each student has a defined mark associated with them so the marks are simply collected together to make the data set.

b) Descriptive statistics : It is the part of statistics that deals with presenting the data we have. This can take two basic forms presenting aspects of the data tither visually with graphs, charts, etc. or numerically and so on.

The basic aim of descriptive statistics is to ‘present the data’ in an understandable way. If you simply write down every piece of data, it means little to someone who sees it, it needs to be summarised.

c) Inferential Statistics : It is the aspect that deals with making conclusions about the data. It used to make inference and describe about the population. These stat are more useful when its not easy or possible to examine each member of the population.