STATISTICS WORKSHEET 1

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

Ans. True

Q2 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

Q3 Which of the following is incorrect with respect to use of Poisson distribution?

Ans Modeling bounded count data

Q4 Point out the correct statement

Ans All of the mentioned

Q5 \_\_\_\_\_\_ random variables are used to model rates.

Ans Poisson

Q6 10. Usually replacing the standard error by its estimated value does change the CLT

Ans False

Q7 1. Which of the following testing is concerned with making decisions using data?

Ans Hypothesis

Q8 4. Normalized data are centred at\_\_\_\_\_\_and have units equal to standard deviations of the original data

Ans 0

Q9 Which of the following statement is incorrect with respect to outliers?

Ans Outliers cannot conform to the regression relationship

Q10. What do you understand by the term Normal Distribution?

Ans A normal distribution has a probability distribution that is centred around the mean. This means that the distribution has more data around the mean. The data distribution decreases as you move away from the centre and the resulting curve is asymmetrical about the mean and form a bell-shaped distribution.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans missing data can occur due to many reasons. The data is collected from various sources and while mining the data, there is a chance to lose the data. Sometimes due to the negligence of people the missing data can be occurred. After classified the patterns in missing values, it needs to be treated like

* Deletion
* Listwise deletion
* Pairwise deletion
* Dropping complete columns

There are several imputation techniques like-

* IMPUTATION WITH CONSTANT VALUE- it replaces the missing value with either zero or any constant value.
* IMPUTATION USING STATISTICS-

1. MEAN-
2. MODE
3. MEDAIN

Q12. What is A/B testing?

Ans A/B testing is also known as split testing or bucket testing. It is a method of comparing two versions of a web page or app against each other to determine which performs better. There are several benefits of this testing like it let you increase user engagement, reduce bounce rates, increase conversion rates, minimum risk, and effectivity create content.

Q13. Is mean imputation of missing data acceptable practice?

Ans No, mean imputation of missing data is not acceptable practice because it ignores feature correlation. The mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate .

Q14. What is linear regression in statistics?

Ans In statistics, linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variable should be quantitative. Linear regression equation **- Y=a+bx+e** where, x is independent variable, B= coefficient value or slope value, A is intercept , Y= dependent variable. We have fit method for training and predict method for testing

15. What are the various branches of statistics?

Ans There are two types of branches of statistics.

* DESCRIPTIVE STATISTICS
* INFERENTIAL STATISTICS

DESCRIPTIVE STATISTICS – descriptive statistics are procedure used to summarize, organize and make sense of a set of scores or observation. Descriptive statistics further divided into two parts

* Central tendency- the four measure of central tendency are **mean**, **mode** and **median**
* Mean= mean is the sum of terms over number of terms
* Mode= mode is the most frequent value.
* Median= median is the mid value of given data
* Dispersion of data – Statistical dispersion means the extent to which numerical data is likely to vary about an average value. In other word dispersion of data means the distribution of data. Dispersion of data is further divided into 4 parts
* Range – range refers to the difference between the largest and smallest value in the given dataset. The higher the value of the range, the higher the spread in data
* Variance – it is the average of the sum of the square of the difference between each data point from the mean. The higher the mean, the higher the scattering of the data from the mean and vice-versa.
* Standard deviation – it is the square root of the variance
* Mean variance – mean variance measures the deviation of data from its central point (MEAN, MEDIAN OR MODE)

INFERENTIAL STATISTICS – inferential statistics helps study a sample of data and make conclusion about its population. A sample is a smaller dataset drawn from the large dataset called the population.