STATISTICS WORKSHEET-1

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

a) True

b) False

Ans : 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

Ans : 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

Ans : b) Modelling bounded count data

4. Point out the correct statement.

a) The exponent of 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 a chi-squared distribution

d) All of the mentioned

Ans **: c) The square of a standard normal random variable follows what is called a chi-squared distribution**

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans : c) Poisson

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

a) True

b) False

Ans : **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

Ans b) Hypothesis

8. Normalized data are cantered at_____ and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

Ans : 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

Ans : c) Outliers cannot conform to the regression relationship

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

Ans : Normal Distributions

Data are symmetrically distributed around the mean, median, and mode. It is a bell-shaped curve distribution. Sometimes it is called the Gaussian distribution.

The form of a normal distribution is determined by it's mean and standard deviation how

Mean = 0 and Standard Deviation = 1

Normal Distribution follows the Three Sigma rule Which is also called the Empirical Rule .

1. *The three sigma rule indicates that given a normal distribution, 68 % of your observations will fall between one standard deviation of the mean.*
2. *95 % will fall within two Sigma and 99.7% will fall within 3 Sigma*
3. *Less than 5 % scores are far from the mean (Not normal score)*
4. *5 % is divided into 2.5 % left and 2.5% right side which is also called the confidence interval*
5. *Confidence interval = 1-95%*

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

Ans : Handling Missing Value

Some common types are

1 Missing completely at Random

2) Missing at Random

3) Missing Not at Random

Technique of dealing with missing values are

1. *Drop missing value /columns/row*
2. *Imputation*

1 *Dropping Missing Data :*
The simplest way to drop columns/rows for which the data is not available above 40 %
But be careful full we can not lose the data set. But a huge number of missing values now apply this method that drops columns/rows.

Dropping the Missing value is not good for any model as the number of data reduces so let us look at a better approach for dealing with missing data.

2 Imputation :

Imputation means to replace or fill the missing data with some value.

There are a lot of ways to impute the data

1 A constant value that belongs to the set of possible values of that variable, such as 0, distinct from all other values

3 A mean, Median, or mode value for the columns

The important point is We all know three types of data int, float and object.

When we impute the missing value in int and float we can use all Mean, Median and Mode But

You can always use Mean and Median but if the data set contains outlier then you use the median as the mean affected by an outlier so better in int and float you must use the Median

12. What is A/B testing?

Ans :

Statistical Analysis to determine if the differences observed between versions A and B are statistically.

A /B testing is also known as split testing .

13. Is mean imputation of missing data acceptable practice?

Ans : *Mean Imputation is a straightforward technique essay to use. It has some drawbacks*

Introduction in Bais in the data set

Distortion of relationship

Distortion in normality

It also impacts in statistically

14. What is linear regression in statistics?

Ans : Linear Regression uses one independent variable to explain or predict the outcome of the dependent variable Y .

Linear Regression Y = a + bx + e

Where Y = The variable that we are trying to predict (Dependent Variable)

X = The input Variable used to predict Y (independent variable)

a = Intercept

b = Slope (Coefficient of x Which is added into Y)

e = Error (regression residual error or cost function)

Linear Regression statistic

Mean Absolute Error =

Mean Square Error =

Root Mean Squared Error =

15. What are the various branches of statistics?

Ans : “Statistics a branch of mathematics used to summarise & analyse and interrupt a group of data or numbers and observations”

Type of Statistics

1 Descriptive stats 2 Inferential stats

1 Measure of Centre	2 Measure of Dispersion	Hypothesis Test
Mean	Range	Zscore
Median	Percentile	T- Test
Mode	IQR	Regression Test
	Standard Deviation	**Chi-square test**
	Variance	ANOVA test
	Skewness	**MNOVA test**
	Kurtosis	**MINCOVA Test**