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) A

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

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

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) D

5. _______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

ANS) C

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

a) True b) False

ANS) B

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

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

ANS) A

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

WORKSHEET

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 or Gaussian distribution refers to a probability distribution where the values of a random variable are distributed symmetrically. These values are equally distributed on the left and the right side of the central tendency. Thus, a bell-shaped curve is formed.*

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

ANS) **Identifying Missing Data**

Missing data occurs in different formats. This section explains the different types of missing data and how to identify them.

Types of missing data

There are three main types of missing data: (1) Missing Completely at Random (MCAR), (2) Missing at Random (MAR), and (3) Missing Not at Random (MNAR).

It is important to have a better understanding of each one for choosing the appropriate methods to handle them.

1) MCAR - Missing completely at random

This happens if all the variables and observations have the same probability of being missing. Imagine providing a child with Lego of different colors to build a house. Each Lego represents a piece of information, like shape and color. The child might lose some Legos during the game. These lost legos represent missing information, just like when they can’t remember the shape or the color of the Lego they had. That information was lost randomly, but they do not change the information the child has on the other Legos.

2) MAR - Missing at random

For MAR, the probability of the value being missing is related to the value of the variable or other variables in the dataset. This means that not all the observations and variables have the same chance of being missing. An example of MAR is a survey in the Data community where data scientists who do not frequently upgrade their skills are more likely not to be aware of new state-of-the-art algorithms or technologies, hence skipping certain questions. The missing data, in this case, is related to how frequently the data scientist upskills.

3) MNAR- Missing not at random

MNAR is considered to be the most difficult scenario among the three types of missing data. It is applied when neither MAR nor MCAR apply. In this situation, the probability of being missing is completely different for different values of the same variable, and these reasons can be unknown to us. An example of MNAR is a survey about married couples. Couples with a bad relationship might not want to answer certain questions as they might feel embarrassed to do so.

Methods for identifying missing data

There are multiple methods that can be used to identify missing data in pandas. Below are the most recurrent ones.

Functions	Descriptions
.isnull()	This function returns a pandas dataframe, where each value is a boolean value True if the value is missing, False otherwise.
.notnull()	Similarly to the previous function, the values for this one are False if either NaN or None value is detected.
.info()	This function generates three main columns, including the “Non-Null Count” which shows the number of non-missing values for each column.
.isna()	This one is similar to isnull and notnull. However it shows True only when the missing value is NaN type.

7 ways to handle missing values in the dataset:

1. Deleting Rows with missing values
2. Impute missing values for continuous variable
3. Impute missing values for categorical variable
4. Other Imputation Methods
5. Using Algorithms that support missing values
6. Prediction of missing values
7. Imputation using Deep Learning Library — Datawig

12. What is A/B testing?

ANS) A/B testing (also known as split testing or bucket testing) is a methodology for comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal .

13. Is mean imputation of missing data acceptable practice?

ANS) The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics? 15. What are the various branches of statistics?

ANS) linear regression- It's a data analysis technique that **predicts the value of unknown information** through the use of another data value.

In other words, this analysis is used to **predict the value of one variable** based on the value of another variable. The variable you want to predict is called the dependent variable, while the second one is called the independent variable.

The method takes into account the coefficients of the **linear equation** and involves different variables that can best predict the value of the variable that is dependent.

In this sense, linear regression **conforms to a straight line** that reduces the differences between the predicted and actual output values. For this purpose, there are now linear regression calculators that use the "least squares" model to choose the line that best fits a set of paired data**.**

15. What are the various branches of statistics?

ANS) **Types of Statistics in Maths**

Statistics have majorly categorised into two types:

1. Descriptive statistics
2. Inferential statistics

Descriptive Statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.