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

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-B

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-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- 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. In graph form, normal distribution will appear as a bell curve.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans- Missing data is a huge problem for data analysis because it distorts findings. It's difficult to be fully confident in the insights when you know that some entries are missing values. Hence, why they must be addressed. According to data scientists, there are three types of missing data. These are Missing Completely at Random (MCAR) — when data is completely missing at random across the
dataset with no discernable pattern. There is also Missing At Random (MAR) — when data is not missing randomly, but only within sub-samples of data. Finally, there is Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing.

Imputation Using (Mean/Median) Values:

This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.

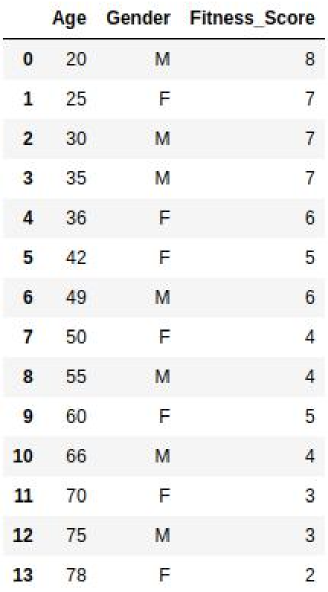
12. What is A/B testing?

Ans- **A/B testing** (also known as **bucket testing** or **split-run testing**) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

13. Is mean imputation of missing data acceptable practice?

Ans-The quick and easy workaround is to substitute a mean for numerical features and use a mode for categorical ones. Even better, someone might just insert 0's or discard the data and proceed to the training of the model. In the following article, I will explain why using a mean or mode can significantly reduce the model's accuracy and bias the results.

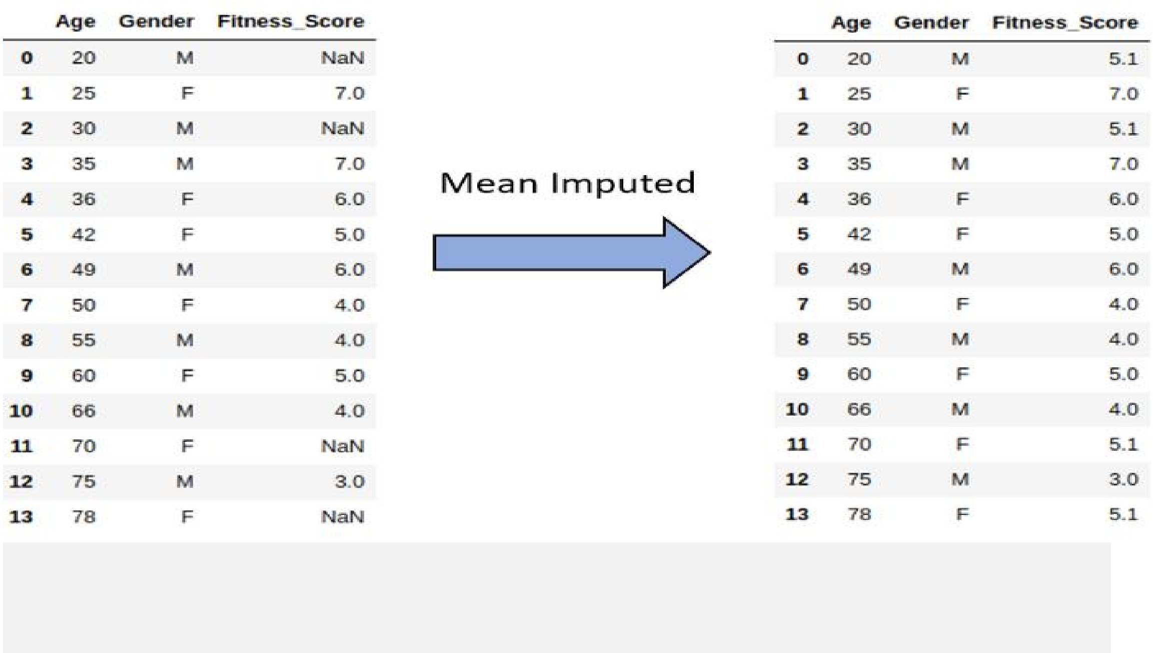
Let's have a look at a very simple example to visualize the problem. The following table have 3 variables: Age, Gender and Fitness Score. It shows a Fitness Score results (0–10) performed by people of different age and gender.



	Age	Gender	Fitness_Score
0	20	M	8
1	25	F	7
2	30	M	7
3	35	M	7
4	36	F	6
5	42	F	5
6	49	M	6
7	50	F	4
8	55	M	4
9	60	F	5
10	66	M	4
11	70	F	3
12	75	M	3
13	78	F	2

Table with correct, non-missing data

Now let's assume that some of the data in Fitness Score is actually missing, so that after using a mean imputation we can compare results using both tables.



	Age	Gender	Fitness_Score
0	20	M	NaN
1	25	F	7.0
2	30	M	NaN
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	NaN
12	75	M	3.0
13	78	F	NaN

Mean Imputed

	Age	Gender	Fitness_Score
0	20	M	5.1
1	25	F	7.0
2	30	M	5.1
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	5.1
12	75	M	3.0
13	78	F	5.1

Mean Imputation of the Fitness_Score

Imputed values don’t really make sense — in fact, they can have a negative effect on accuracy when training our ML model. For example, 78 year old women now has a Fitness Score of 5.1, which is typical for people aged between 42 and 60 years old. **Mean imputation doesn’t take into account a fact that Fitness Score is correlated to Age and Gender features.** It only inserts 5.1, a mean of the Fitness Score, while ignoring potential feature correlations.

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 (also known as dependent and independent variables). The case of one explanatory variable is called *simple linear regression*; for more than one, the process is called **multiple linear regression**. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

15. What are the various branches of statistics?

Ans- – There are two main branches of statistics –

1- Inferential Statistic.

2- Descriptive Statistic.

Inferential Statistics: Inferential statistics used to make inference and describe about the population. These stats are more useful when it’ s not easy or possible to examine each member of the population.

Descriptive Statistics: Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.