

## WORKSHEET 1

### Statistics

Q1. Ans- A] True

Q2. Ans- A] Central limit theorem

Q3. Ans- C] Modeling bounded count data

Q4. Ans- D] All the mentioned

Q5. Ans- C] Poisson

Q6. Ans- B] False

Q7. Ans- B] hypothesis

Q8. Ans- A] 0

Q9. Ans- C] Outliers cannot conform to the regression relationship

Q10. 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.

Q11. Ans-

Best techniques to handle missing data :-

1. Use deletion method to eliminate missing data.
2. Use regression analysis to systematically eliminate data.
3. Data scientist can be use data imputation techniques.
4. Keeping things under control.
5. Assigning a unique value.
6. Predicting a missing value.

The best method is to deletion method to eliminate missing data as it ensures that no bias or variance is added or removed, and ultimately results in a robust and accurate model. However, this is only recommended if there's a lot of data to start with and the percentage of missing values is low.

Q12. Ans-

1. A/B testing is a popular way to test your products and is gaining steam in the data science field.
2. A/B testing 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.
3. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

Q13. Ans-

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.

Q14. Ans-

Linear regression is an attempt to model the relationship between two variables by fitting a linear equation to observed data, where one variable is considered to be an explanatory variable and the other as a dependent variable.

Q15. Ans- Statistics have two main branches, namely:

- 1] Descriptive Statistics: This usually summarizes the data from the sample by making use of an index like mean or standard deviation. The methods which are used in the descriptive statistics are displaying, organizing, and describing the data.
- 2] Inferential Statistics: These conclude from data which are subject to random variations like observation mistakes and other sample variation.