STATISTICS WORKSHEET-1

ANSWERS

Q1> ANS: (a) **True**

Q2> ANS: **(a) Central Limit Theorem**

Q3> ANS: **(b) Modelling bounded count data**

Q4> ANS: **(d) All of 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**

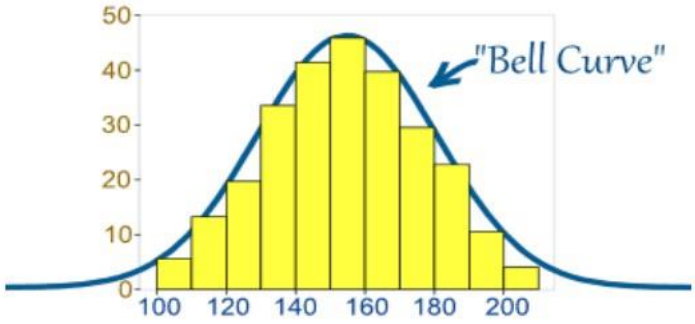
Q10.**What do you understand by the term Normal Distribution?**

- The normal distribution is the most widely known and used of all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems.

  The normal distribution, also known as the Gaussian distribution, is the most important probability

distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

approximately normally distributed; measurement errors also often have a normal distribution. The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal.



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

The problem of missing value is quite common in many real-life datasets. Missing value can bias the results of the machine learning models and/or reduce the accuracy of the model

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset.
Below is a sample of the missing data from the Titanic dataset. You can see the columns ‘Age’ and ‘Cabin’ have some missing values.

Reasons for the missing data from the dataset affect the approach of handling missing data. So it's necessary to understand why the data could be missing.

Some of the reasons are listed below:

- Past data might get corrupted due to improper maintenance.
- Observations are not recorded for certain fields due to some reasons. There might be a failure in recording the values due to human error.
- The user has not provided the values intentionally.

first step in handling missing values is to look at the data carefully and find out all the missing values.

The following code shows the total number of missing values in each column.

It also shows the total number of missing values in entire data set.

Real-world data is messy and usually holds a lot of missing values. Missing data can skew anything for data scientists and, A data scientist doesn't want to design biased estimates that point to invalid results. Behind, any analysis is only as great as the data. **Missing data appear when no value is available in one or more variables of an individual**

There are different ways of replacing the missing values.

Missing Completely at Random(MCAR):

A variable is missing completely at random (MCAR)if the missing values on a given variable (Y) don't have a relationship with other variables in a given data set or with the variable (Y) itself. In other words, When data is MCAR, there is no relationship between the data missing and any values, and there is no particular reason for the missing values.

Missing Not at Random(MNAR):

The final and most difficult situation of missingness. MNAR occurs when the missingness is not random, and there is a systematic relationship between missing value, observed value, and missing itself. To make sure, If the missingness is in 2 or more variables holding the same pattern, you can sort the data with one variable and visualize it.

Imputation techniques:

The imputation technique replaces missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the data and its problem.

Imputation using Statistics:

The syntax is the same as imputation with constant only the SimpleImputer strategy will change. It can be "Mean" or "Median" or "Most_Frequent".

“Mean” will replace missing values using the mean in each column. It is preferred if data is numeric and not skewed.

“Median” will replace missing values using the median in each column. It is preferred if data is numeric and skewed.

“Most_frequent” will replace missing values using the most_frequent in each column. It is preferred if data is a string(object) or numeric.

Before using any strategy, the foremost step is to check the type of data and distribution of features(if numeric).

Q12. **What is A/B testing?**

- A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.

  The concept is similar to the scientific method. If you want to find out what happens when you change one thing, you have to create a situation where only that one thing changes.

Q13.**Is mean imputation of missing data acceptable practice?**

- 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.

Q14. What is linear regression in statistics?

Linear Regression:

- Linear regression analysis is used to predict the value of a variable based on the value of another variable

If we want to use a variable x to draw conclusions concerning a variable y: y is called dependent or response variable. x is called independent, predictor, or explanatory variable. If the relationship between two variables is linear is can be summarized by a straight line. A straight line can be described by an equation:

y= a + bx

Q15.What are the various branches of statistics?

- There are basically four branches into which statistics is divided.

1. **<u>Mathematical or theoretical statistics</u>**

   It helps in forming the experimental and statistical distribution.

2. Statistical methods or functions

It helps in the collection, tabulation and interpretation of the data. It helps in analyzing the data and returns insight from the data

3. Descriptive statistics

It helps in summarizing and organizing any data set characteristics. It also helps in the representation of data in both classification and diagrammatic way.

4. Inferential Statistics

Inferential statistics are often used to compare the differences between the treatment groups. Inferential statistics use measurements from the sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects.