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

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

3) Which of the following is incorrect with respect to use of Poisson distribution?
   d) All of the mentioned

4)Point out the correct statement.
   d) All of the mentioned

5) ______ random variables are used to model rates.
  c) Poisson

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

7) Which of the following testing is concerned with making decisions using data?
   b) Hypothesis

8) Normalized data are centered at______and have units equal to standard deviations of the original data.
   a) 0

9) Which of the following statement is incorrect with respect to outliers?
   c) Outliers cannot conform to the regression relationship

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?

In probability theory and statistics, the Normal Distribution, also called the Gaussian Distribution, is the most significant continuous probability distribution. Sometimes it is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. Furthermore, it can be used to approximate other probability distributions, therefore supporting the usage of the word 'normal 'as in about the one, mostly used.

Normal Distribution is defined by the probability density function for a continuous random variable in a system.

Normal Distribution Formula

The probability density function of normal or gaussian distribution is given by;

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where,

- x is the variable
- μ is the mean
- σ is the standard deviation

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

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset.
In the dataset, blank shows the missing values.
In Pandas, usually, missing values are represented by NaN.
It stands for Not a Number.
There can be multiple reasons why certain values are missing from the data.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.

How To Handle The Missing Data

Analyze each column with missing values carefully to understand the reasons behind the missing values as it is crucial to find out the strategy for handling the missing values.

There are 2 primary ways of handling missing values:

1. Deleting the Missing values
2. Imputing the Missing Values

Deleting the Missing value

Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values.If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted.If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted.

Imputing the Missing Value

There are different ways of replacing the missing values. You can use the python libraries Pandas and Sci-kit learn as follows:

Replacing With Arbitrary Value

Replacing With Mean

Replacing With Mode

Replacing With Median

Replacing with previous value – Forward fill

Replacing with next value – Backward fill

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?

A/B testing (also known as split testing or bucket testing) is a method of 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.

Running an A/B test that directly compares a variation against a current experience lets you ask focused questions about changes to your website or app and then collect data about the impact of that change.

Testing takes the guesswork out of website optimization and enables data-informed decisions that shift business conversations from "we think" to "we know." By measuring the impact that changes have on your metrics, you can ensure that every change produces positive results.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is a method of replacing missing data with the mean value of the available data. While it can be an acceptable method in some cases, it is important to consider the potential consequences of using this method and to ensure that it is appropriate for the specific dataset and analysis being conducted.

There are a few potential problems with using mean imputation:

It can introduce bias into the data if the missing values are not randomly distributed. For example, if the missing values are more likely to occur for certain groups or subpopulations within the dataset, then using the mean value to impute the missing data could lead to biased results.

It can reduce the variance of the data, which can impact the statistical power of the analysis.

It can be inappropriate if the data are not normally distributed, as the mean may not accurately represent the central tendency of the data.

Therefore, it is important to carefully consider the potential impacts of mean imputation on the dataset and analysis, and to choose an appropriate method for dealing with missing data based on the specific circumstances. Other methods for handling missing data, such as multiple imputation or excluding observations with missing values, may be more appropriate in certain cases.

14. What is linear regression in statistics?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15. What are the various branches of statistics?

The two main branches of statistics are

1)Descriptive statistics and

2) Inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.
Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.
Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics.
Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.