Worksheet Set: 1

Statistics

1. Bernoulli random variables take (only) the values 1 and 0.

= True (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?

= Central Limit Theorem. (A)

3. Which of the following is incorrect with respect to use of Poisson distribution

= Modeling Bounded Count Data. (B)

4. Point out the correct statement.

= All of the mentioned. (D)

5. _____ random variables are used to model rates.

= Poisson (C)

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

= False. (B)

7. Which of the following testing is concerned with making decisions using data?

= Hypothesis. (B)

8. Normalized data are centered at _____ and have units equal to standard deviations of the original data.

= 0. (A)

9. Which of the following statement is incorrect with respect to outliers?

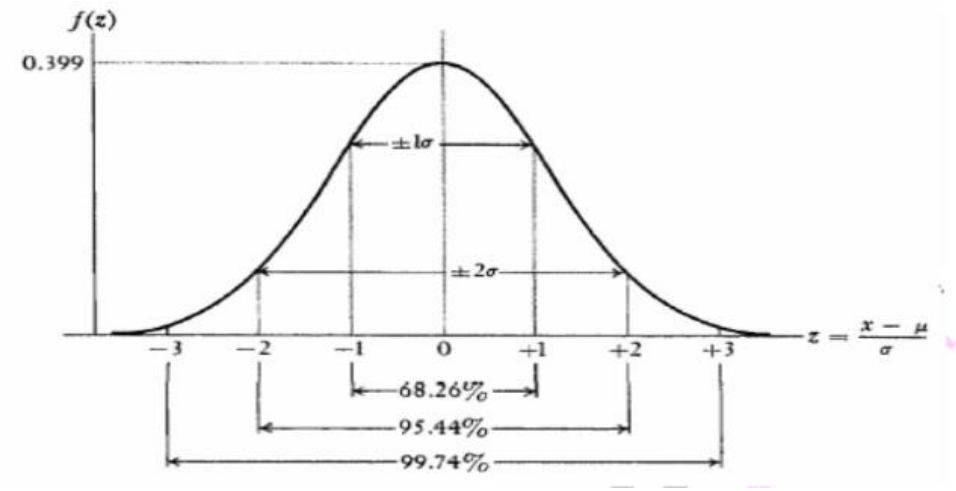
= Outliers can't confirm to the regression relationship. (C)

10. What do you understand by the term Normal Distribution?

= If the diagram looks something like a bell shape i.e.

Mean = Median = Mode

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



z = (x - μ)/Std. deviation.

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

= When we get the dataset from the source, there is a high chance that we will be getting the dataset having missing data or null values. And we must clean the dataset before training the model. So, first we have to check for the missing data or null values using pandas library.

syntax:

import pandas as pd

dataset.isna().sum()

This syntax will show the count of missing data that a particular column has.

There are 2 ways to handle the missing data:

1. **Deleting the missing values.**

   In deleting the missing values, we used to delete the entire row or column as per our analysis.

   syntax: data.dropna()

2. **Imputing the missing values**

   In imputing the missing values, we do not delete anything from the dataset instead we give them some data or fill the values in the missing values. This is done by the imputation techniques; we are having so many imputation techniques through which we can give the values to the missing values.

Some of the Imputation techniques are:

- **Filling all the null values with 0.**

  syntax: dataset['column_name'].fillna(0)

- **Filling all the null values with the mean/median/mode of the column.**

  syntax:
  dataset['column_name'].fillna(dataset['column_name'].mean()/median()/mode())

- **Replacing with forward value or backward value.**

  syntax: dataset['column_name'].fillna(method='ffill'/'bfill')

- **By using the Simple Imputer/knn imputer.**

  syntax:

  from sklearn.impute import SimpleImputer,KNNImputer

  si = SimpleImputer()

  si.fit_tranform(dataset['column_name'])

  knn = KNNImputer(n_neighbors = n)

  knn.fit_transform(dataset['column_name'])

12. What is A/B testing?

= A/B testing is basically statistical hypothesis testing. It is an analytical method for making decisions that estimates population parameters based on sample statistics.

The A/B testing process can be simplified as:

- **A/B testing starts with process by making hypothesis.**
- **The test to gather statistical evidence to accept or reject a hypothesis.**
- **The final data shows whether the hypothesis is correct or incorrect.**

Hypothesis can also be termed as impression, making a guess based on assumption without scientific proof or explaining the situation based in reasonable assumption.

- **Null Hypothesis (ho)**

  Decision always leads to assumption doesn't change.

- **Alternate Hypothesis (ha)**

  Decision leads to opposite of null hypothesis (ho)

13. Is mean imputation of missing data acceptable practice?

= No mean imputation of missing data is not an acceptable practice, as it will fill all the missing data with the mean and that is not acceptable for example if we have the dataset of cars and having the missing values in it, there is a chance that will be end up with the non-explainable dataset. Suppose our data set is having data of cars: BMW, Audi, Honda and we are having the missing value in it, and we use mean imputation in the dataset, then it can be result as the false value or incorrect data.

So using mean imputation can put us in trouble in some of the cases.

14. What is linear regression in statistics?

= Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data

points and plots a trend line. Linear regression can create a predictive model on apparently random data, showing trends in data.

Linear regression shows a relationship between an independent variable and a dependent variable.

15. What are the various branches of statistics?

= Branches of statistics:

- **Descriptive statistics**

  If data can be described without any statistical tools, then it is called descriptive statistics.

  - **It summarizes or describes the characteristics of a data set.**
  - **Descriptive statistics consists of two basic categories of measures: measures of central tendency and measures of variability**
  - **Measures of central tendency describe the center of a data set.**
  - **Measures of variability or spread describe the dispersion of data within the set.**

  For example, Marks in class, Height of student.

- **Inferential statistic**

  If the data is too big, then we use inferential statistics. We take a few samples from different data, and we find the average. This is called inferential statistics. The average is then applicable to all the data from where we have selected our samples.

  - **Inferential statistics can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples.**
  - **The goal of inferential statistics is to make generalizations about a population.**
  - **In inferential statistics, a statistic is taken from the sample data (e.g., the sample mean) that used to make inferences about the population parameter (e.g., the population mean).**