STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
Ans: TRUE (Bernoulli Random Variables works on Binomial Distribution. It works on 0 and 1.

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?

Ans: Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
Ans: Modeling event/time data

4. Point out the correct statement.
Ans: All three statements are right
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

5._____ random variables are used to model rates.
Ans: c) Poisson

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

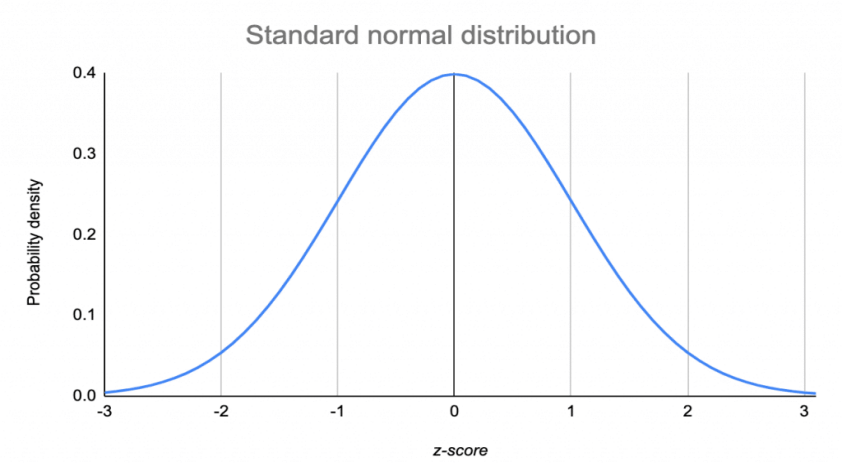
7. Which of the following testing is concerned with making decisions using data?
Ans: b) Hypothesis

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

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

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: A normal distribution also called Gaussian distribution of data is one in which the majority of data points are relatively similar, meaning they occur within a small range of values with fewer outliers on the high and low ends of the data range. In such a distribution of data, mean, median, and mode are all the same value and coincide with the peak of the curve.



11. How do you handle missing data? What imputation techniques do you recommend?
Ans: The Best techniques to handle missing data are given below:

- First, determine the pattern of your missing data and Use deletion methods to eliminate missing data
- Use regression analysis to systematically eliminate data
- Data scientists use two data imputation techniques to handle missing data.
- Average imputation and common-point imputation.

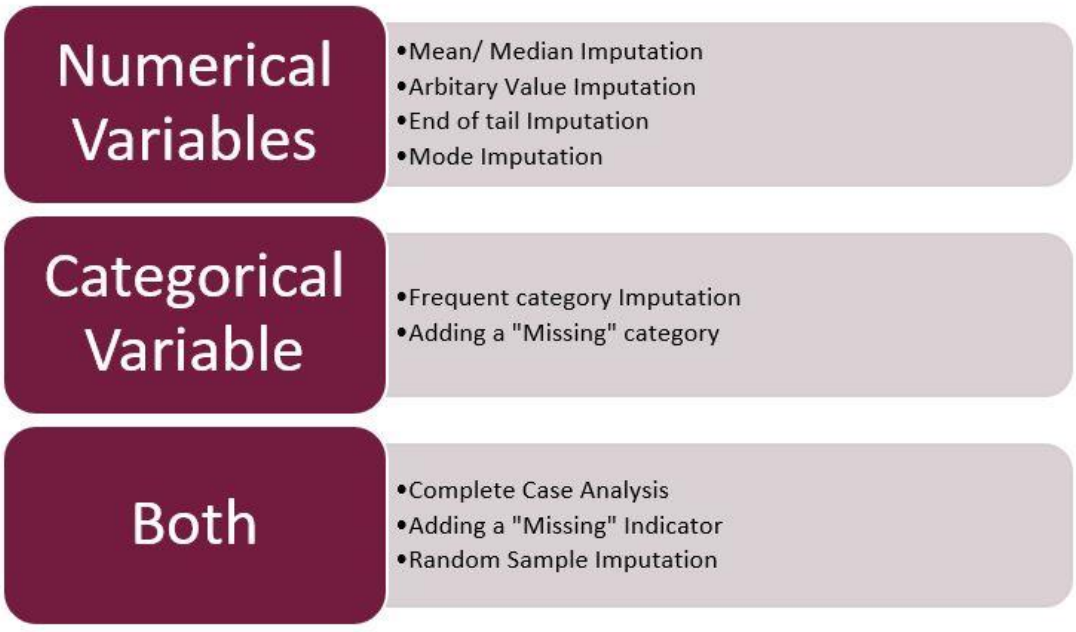
The imputation techniques we recommend to handle missing data:Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. These techniques are used because

removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis.

We use imputation because Missing data can cause the below issues: –

1. Incompatible with most of the Python libraries used in Machine Learning:- Yes, you read it right. While using the libraries for ML (the most common is skLearn), they don't have a provision to automatically handle these missing data and can lead to errors.
2. Distortion in Dataset:- A huge amount of missing data can cause distortions in the variable distribution i.e. it can increase or decrease the value of a particular category in the dataset.
3. Affects the Final Model:- The missing data can cause a bias in the dataset and can lead to a faulty analysis by the model.

Imputation Techniques



12. What is A/B testing?

Ans: 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. It is also known as bucket testing or split- run testing. It includes application of statistical hypothesis testing or two sample hypotheses 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 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.

14. What is linear regression in statistics?

Ans: Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula $y = c + b*x$, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

15. What are the various branches of statistics?
Ans: Statistics is the branch of mathematics that deals with data. It is used to accomplish a variety of tasks, such as data collecting, organization, and analysis. A collection of data is often referred to as a data set or set of data¸ but other words such as a list or simply collection are also often used.
Examples of data sets are:
Marks in a class test: 9, 2, 5, 8, 10, 3, 5, 8, 8, 9
Inflation rate: 2.1, 3.2, 4.1, 2.3, 5.1, 2.2, 0.5
Voting intention in a referendum: Yes, No, No, Yes, Yes, No

Two branches, descriptive statistics and inferential statistics, comprise the field of statistics.

Descriptive Statistics

This branch of statistics that focuses on collecting, summarizing, and presenting a set of data. The first aspect of statistics is descriptive statistics, which deals with the presentation and collecting of data. It is not as simple as it appears, and the statistician must be aware of how to design and experiment, select the appropriate focus group, and prevent biases that are all too easy to introduce into the experiment.

Inferential Statistics

This branch of statistics that analyzes sample data to draw conclusions about a population. Inference statistics are statistical techniques that allow statisticians to utilize data from a sample to conclude, predict the behavior of a given population, and make judgments or decisions. Using descriptive statistics, inference statistics frequently talk in terms of probability.