STATISTICS WORKSHEET- 1

ANSWER: Question 1 to question 9:

1. True
2. Central Limit Theorem
3. Modeling bounded count data
4. All of the mentioned
5. Poisson
6. False
7. Hypothesis
8. 0
9. Outliers cannot conform to the regression relationship

Answer Question 10 to Question 15:

Q:10 Answer- The normal distribution is an important class of Statistical distribution that has a wide range of application. This distribution applies in most Machine learning algorithms and the concepts of the normal distribution is a must for any Statistician, Machine Learning Engineer, and Data scientist.

The normal distribution is also known as Gaussian or Gauss distribution. Many group follow this type of pattern. That’s why its widely used in business, statistics, and in government bodies like that FDA:

1. Heights of people
2. Measurement errors.
3. Blood pressure
4. Points on a test
5. IQ scores
6. Salaries

Question 11 Answer:

Handle missing data is an inevitable part of the process. As data researchers, we pour a lot of resources, time and energy into making sure the data set is as accurate as possible. However, data inevitable goes missing. As someone who has been handling data analytics and overseen dozen of research project for several years, missing data is just one of those it sucks but its no one fault scenarios. Sometimes data sets come up short no matter how mant times data scientist clean and prepare it. The best way to handle such situations is to develop contingency plans to minimize the damage.

1. Use deletion methods to eliminate missing data
2. Use regression analysis to systematically eliminate data
3. Data scientists can use data imputation techniques

Imputation techniques are recommend:

1. Complete Case Analysis (CAA): This is quite straight forward method of handling the missing data which directly removes the rows that have missing data we consider only those rows where we have complete data and data is not missing. This method is also known list wise deletion:

Assumption:

1. Data is Missing At Random (MAR)
2. Missing data is completely removed from the table

Advantages:

1. Easy to implement
2. No data manipulation required

Limitation:

1. Deleted data can be informative.
2. Can lead to the deletion of a large part of the data
3. Can create bias in the dataset, if large amount of a particular type of variables is deleted from it.
4. The production model will not know to do with missing data.

When to Use:

1. Data is MAR.
2. Good for Mixed, Numerical and Categorical data.
3. Missing data is not more than 5%-6% of the dataset
4. Data does not contain much information and will not bias the dataset.

2.) Arbitrary Value Imputation:

This is an important techniques as it can handle both the numerical and categorical variables. This techniques states that we group the missing values in a column and assign them to a new value that is far away from the range if that column. Mostly we use values like 999999 or 9999999 or Missing or Not defined for numerical and categorical variables.

Assumption:

1. Data is not MAR.
2. The missing data is imputed with an arbitrary value that is not part of the dataset pr Mean/Median/Mode of data.

Advantages:

1. Easy to implement.
2. We can use it in production.
3. It retained the importance of missing values if it exits.

Disadvantages:

1. Can distort original variable distribution.
2. Arbitrary values can create outliers.
3. Extra caution required in selecting the Arbitrary value.

When to Use:

1. When data is not MAR.
2. Suitable for all.

3.) Frequent Category Imputation:

This techniques says to replace the missing value with the variables with the highest frequency or in simple words replacing the values with the mode of that column. This techniques is also referred to as Mode Imputation.

Assumption:

1. Data is MAR.
2. There is a high probability that the missing data looks like the majority of the data.

Advantages:

1. Implementation is easy.
2. We can obtain a complete dataset in very little time.
3. We can use this techniques in the production model.

Disadvantages:

1. The higher the percentage of missing values, the higher will be distortion.
2. May lead to over representation of a particular category.
3. Can distort original variable distribution.

When to Use:

1. Data is MAR.
2. Missing data is not more than 5%-6% of the dataset.

Question 12 Answer:

The A/B Testing in its simplest sense is an experiment on two variants to see which perform better based on a given metric. Typically to consumer groups are exposed to two different versions of the same thing to see of these is a significance difference in metrics like sessions click through rate and conservations.

Using the visual above as an example we could randomly split our consumer based into two groups a control group and a variant group. Then we can expose our variant group with a red website banner and see if we get a significant increases in conservation. Its is important to note that all other variables needs to be held constant when performing an A/B Testing.

Getting more technical, A/B testing is a form of statistical and two sample hypothesis testing. Statistical hypothesis testing is a method in which a sample database is composed against the population data. Two sample hypothesis testing is a method in determined whether the differences between two samples are statistically or not.

Question 13 Answer:

Mean imputation of missing data acceptable practice are:

1. Bad practical in general.
2. If just estimated means: mean imputation preserves the mean of the observed data.
3. Leads to an underestimate of the standard deviation.
4. Distorts relationship between variables by pulling estimates of the correlation toward zero.

Question 14 Answer:

The 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 predicating an outcome variable.
2. Which variables in particular are significant predictors of the outcomes variable and in what way do they indicated by the magnitude and sign of the beta estimates impact the outcome variables. 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 = estimate dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variables.

Three major uses for regression analysis are;

1. Determining the strength of predictors.
2. Forecasting an effect.
3. Trend forecasting.

Question 15 Answer:

The various branch of statistics are:

1. Descriptive Statistics: It describes the important characteristics properties of the data using the measures the central tendency like mean/median/mode and the measures of dispersion like range, standard deviation variance etc.

Data can be summarized and represented in an accurancy way using charts, tables and graphs.

For example: We have marks 1000 students and we may be interested in the overall performance of those students and the distribution as well the spread of marks. Descriptive statistics provides us the tools to define our data in a most understandable and appropriate way.

1. Inferential Statistics:

It is about using data from sample and the making inferences about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalized them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, analysis of variance etc.

For example: Suppose we are interested in the exam marks of all the student in India. But it is not feasible to measure the exam marks of all the students in India. So now we will measures the marks of a smaller sample of students for example 1000 students. This sample will now represent the large population of Indian students. We would consider this sample for our statistical study for studying the population from which its deducted.

THE END