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?
b) Modeling bounded count data

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

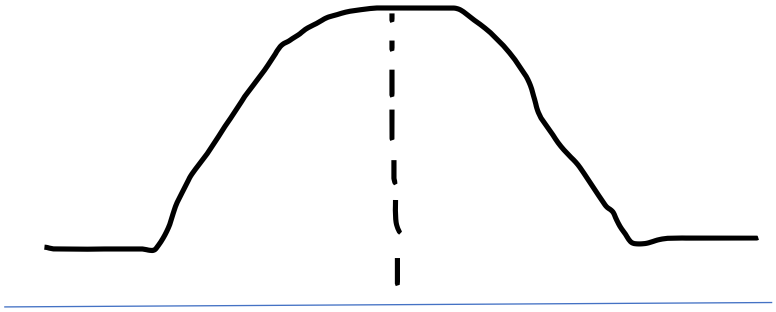
9. Which of the following statement is incorrect with respect to outliers?
c) Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

A) A Normal Distribution is a type of probability distribution of random variables in which the values are symmetrically distributed. In this type of distribution Mean, Median and Mode are equal or very close to each other. It forms a bell-shaped curve in statistics graphs and reports. Most of the data values are around the mean. The value of standard deviation is 1 or very close to 1 in some instances. 99% of data lies within the range of 1 to 3 standard deviations. It has got skewness equal to 0 and kurtosis equal 3. There is constant proportion of distance lying under the curve.

It is graphically represented as:



Normal distributed data is most preferred distribution for model building in machine learning.

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

A: In any given machine learning problem first and foremost important input is data. When we missing values in data our model or findings will never attain high accuracy and even the output will be troublesome resulting in losses. This happens mainly due to data inconsistency. Therefore, to avoid this problem we use different methods depending on our feasibility. Manually if we have resources then we can either re-check the data and fill the missing values or redo the survey and fill the dataset more accurately to avoid human error.

If we do not have accessibility to them, we can even do using python packages. Techniques to handle missing values:

1st. **Fillna**: we can fill missing values using specific methods like 0, dict or series etc.

2nd. **Dropna**: We can drop the missing values from the dataset

3rd. **Impute**: We can impute missing values by replacing them by most suitable value and maintaining the essence of dataset.

Depending on the dataset we choose handle missing data accordingly. We can drop missing values we they are either less in number or abundant in column then we drop that feature itself. There are different techniques for data imputation we can use from scikit-learn package. They are as follows:

a) Simple Imputer- the most basic method of imputing the data where we can decide upon the strategy to fill missing values either by mean or median or mode.
b) K-NN imputer- Complex method which finds nearest possible value depending on adjacent values
c) Iterative Imputer- it is more sophisticated multivariate approach. In this, missing values is treated as a function of other features to estimate its value.

Personally, I find Iterative imputer to be a better technique for imputation. This technique imputes the value by iterating in a round robin fashion and taking other features into account. It is resource consuming but output is closest to possible value. It helps in retaining the character of the dataset.

In conclusion, we cannot narrow down to a single missing value handling technique. It is entirely based on data and situation we are dealing in. We have to at times use try-error method to find the best possible technique.

12. What is A/B testing?

A: A/B testing is a type of statistical test where we use test of hypothesis to understand the relationship between two data sets and know its statistical significance. In other words, A/B testing is a basic randomized control experiment. In this test we make two hypothesis which are NULL hypothesis (Ho) and ALTERNATE hypothesis (Ha). In Null hypothesis we state that observations result is purely from chance. In Alternate hypothesis we state converse of Null hypothesis that variables play a vital role to influence observations result. After defining our hypothesis, we divide the data set into two equal groups A (the control group) & B (the test group). The sample observations in each group are taken randomly to avoid sampling bias. Moving forward we conduct the test and record results. Our inferences are subject to two types of errors: Type I Error in which our Null Hypothesis is true but we choose Alternate hypothesis & Type II error in which our Null hypothesis is false but we fail to reject it. Depending on the P-value being greater or less than 0.05(level of significance) we decide whether Ho is to be accepted or rejected.
This is how A/B testing helps us in making decisions and is a crucial technique.

13. Is mean imputation of missing data acceptable practice?

A: Mean imputation of a missing data punches in the mean value of the feature column in the missing value. Though it is pretty handy and safe affair, we cannot always rely on this practice. Having said that, usage of techniques to fill-in the missing values depends entirely upon the nature and situation of dataset. Some probable pros and cons of using mean strategy could be:

Pros

- It uses average to fill the missing data
- It is a faster and economical technique

Cons

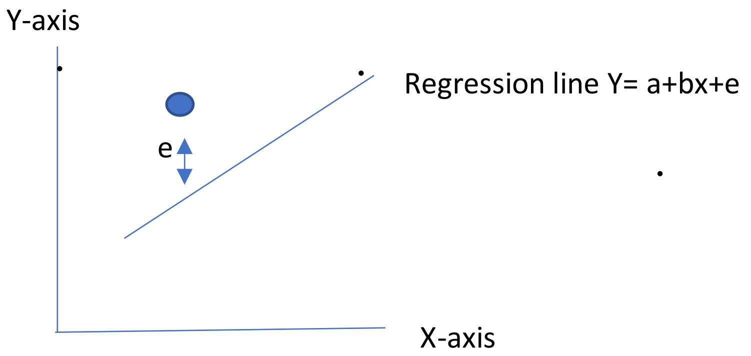
- It will impute data with mean for multiple rows in same column making data inconsistent
- It may change the probable value of missing drastically when compared to other more efficient techniques
- It is a univariate technique which means it does not take other features into account

To conclude, mean imputation can not solely be acceptable technique at times in a dataset. Hence, we need to use other techniques too to get best possible result.

14. What is linear regression in statistics?

A: Linear Regression is a type of statistical model which explains the variation of a dependent variable based on the variation of independent variable/s. It is a type of predictive model. It only deals with continuous data and cannot be used for categorical data. It is a simple but effective model for not complex machine learning problems. The equation for linear regression is:
Y= a + bx + e

Where,
Y- predicted value
a- intercept (mean value of Y when independent variable “x” is 0)
x- given value of an independent variable
b- slope (Rate of change in Y for unit change in “x”)
e- error term



The assumptions for linear regression are:

1st. Linear in variable: The relationship between x and mean of Y is linear in parameters
2nd. Homoscedasticity: The variance of residual or error is the same for any value of x
3rd. Independence: The observations are independent of each other
4th. Normality: For any fixed value of x, Y is normally distributed
5th. Mean of the residual is 0
6th. The number of observations must be greater than number of x
7th. The regression model is correctly specified
8th. The variability in x values must be positive
9th. The feature variables or x must not have multicollinearity

One must keep the assumptions in mind as they are very vital in drafting linear regression model.

15. What are the various branches of statistics?

A: According to Oxford dictionary, Statistics is defined as the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample. Due to advancement in technologies and researches, we use statistics to make important business, national policies and health decisions. Statistics is widely used in all walks of life knowingly or unknowingly. In most basic thought predicting showers before stepping out from home includes probability which is also a part of statistics.

Statistics for our ease of understanding is basically divided into two branches:

1. Descriptive Statistics:
   This branch of statistics deals mainly with describing and presentation of the data. It is the primitive and forms base to understand the data. In this step we get to know about the mean, median, quartiles, standard deviation etc. We understand the behaviour of the data with respect to its centre point. The features and quality of dataset is summarized using descriptive statistics.
2. Inferential Statistics:
   This branch is the second step in our statistical findings. Once we have the data described and have learned about its movement from the centre, we can infer to make a decision. In this branch we use data from the sample to conclude for the population behaviour. It frequently includes probability. Other types of inferential statistics are regression analysis, ANOVA, Correlation analysis etc. We use inferential statistics to test hypothesis or to check if sample data can be generalized for larger population.