Statistics worksheet-1

Q1) The correct answer is - a

True.

Q2) The correct answer is - a

The central limit theorem.

Q3) The correct answer is - b

because poission distribution is used for unbounded count data.

Q4) The correct answer is - d

Q5) The correct answer is - c

Poission distribution is used to model counts.

Q6)The correct answer is - b

usually replacing the standard error by its estimated values does not changes the CLT.

Q7)The correct answer is - b

Hypothesis testing is used for making decision using data.

Q8)The correct answer is - a

normalized data are centered at 0.

Q9) The correct answer is - c

Outliers can conform to the regression relationship.

Q10) **Normal distribution-**

Normal distribution also known as Gaussain distribution or looks like a bell curve.

In a normal distribution the mean is 0 and the spread or the standard deviation is 1 with 0 Skewness.

Normal distributions are symmetrical in shape.

In normal distributions 68% observations will appear from +/-1 std. , 95% observations will appear from +/-2 std. and 99.7% observations appears from +/-3std.

Q11)

Missing data or the nan values can be filled with the help of Fillna method.but if we have have thousands of features, in that case we cant write fillna method for each feature again n again.so to resolve this problem we have different imputation techniques which we generally use in ML.

1. Simple imputers-

Replacing missing values of feature or column using a descriptive statistic (e.g. mean, median, or most frequent) of that particular column.

2.) Knn imputers-

KNN is an algorithm that is useful for matching a point with its closest k neighbors in a multi-dimensional space. It can be used for data that are continuous, discrete, ordinal and categorical which makes it particularly useful for dealing with all kind of missing data.

In this imputation techniques we have to provide the k neighbours values and supporting neighbor column along with the column in which we want to fill missing values.

Suppose if we have given k=2 than this method will fill the nan value with the help of average of two values of the same column, these two values are being found by two nearby values of the supporting column.

3.) Iterative imputers-

This particular type of imputation technique works on linear regression principle.the whole data will be divided into train and test data.

The values which we want to fill will be the test data and remaining data will be our train data.

So by this way this method use to predict the nan values with the help of linear regression and gradient descent.

Q12) A/B Testing-

A/B testing (also known as split testing or bucket testing) is a methodology for 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.

In an A/B test, we take a web page or app screen and modify it to create a second version of the same page. This change can be as simple as a single headline, button or be a complete redesign of the page. Then, half of your traffic is shown the original version of the page (known as control) and half are shown the modified version of the page (the variation).

As visitors are served either the control or variation, their engagement with each experience is measured and collected in a dashboard and analyzed through a statistical engine. You can then determine whether changing the experience (variation) had a positive, negative or neutral effect against the baseline (control).

Q13.) The process of replacing null values in data collection with data mean is known as mean imputation.

Mean imputation is a terrible practice since it doesnot consider feature correlation. Suppose we have two features age and fitness score.an 10 year old boy fitness score is missing. Now with help of mean imputation if we try to fill it with mean of all age group fitness score, which will comeout very high which is comparitively very large for a 10 year old boy.so that’s where this mean imputation technique becomes terrible.

Q14.) Linear regression-

Linear regression is used to predict continous data or numerical data.

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

It works on the principle of gradient descent, which is nothing but tries to reduce error in each iteration and tries to reach global minima with the hyper tunning of learning rate.

Linear regression create the best fit regression line to predict the label or continous data with help of equation of straight line-

Y=mx+c

Where y is actual value (mx+c) is predicted value in which m is the slope or coefficient, c is the intercept,and x is the data value.

Linear regression can be used to find the relationship between features and label and the impact of features on label.

With the help of linear regression model we use to predict the label and we can check the accuracy of the model also and the train and test score also.

To check the overfitting of the model we can use Ridge and Lasso regularization technique.

Q15.) Branches of statistics-

Statistics is something used to deal with the data or the numbers either collection of these numbers or studying these numbers.

1. Descriptive statistics-

Measuring or describing something by measure of central tendency(mean,median,mode) and Variance and std.

Descriptive statistics is used when we can describe something. Suppose if someone ask us which party will win in election in our city.with considering a number of factors we can predict the x party has more chances to win. Because in city the area is limited so we can describe it.

2. Inferential statistics-

Inferential statistics is used when we cant describe something. Suppose rather than asking about the election result prediction of our city someone ask us the election result predict of whole india than it will become very difficult for us to predict the result of whole india because the area is very large.

In that case what we will do is, we will take samples from each states and will take the mean of the samples.

Than the total average of the sample mean will be the mean of the population.

So basically the results we get from samples, we infer them to whole poplutaion.