Inferential statistics

What and why :

As its name suggests, IS is about making inferences about population from the sample. It generalises to large dataset and applies probabilities to draw conclusion. It is mainly used to get interpretation and draw conclusion. IS is mainly related to and associated with hypothesis testing whose main target is to reject null hypothesis. IS is generally used to determine how strong relationship is within sample.

In practical scenario we use IS, like we can not ask 140 crores people about their happiness instead we pick samples and from those samples we generalise our conclusion to whole population. In ML the concept of hypothesis is mainly used. The hypothesis is explained below.

Hypothesis :

In inferential statistics we get insights from sample data and draw conclusion for population data. So this conclusion or assumption is called hypothesis. To draw conclusion for population data we need to do hypothesis testing. Here we make null hypothesis and alternate hypothesis. Null hypothesis is that we define the positive hypothesis. And alternate hypothesis is the opposite of null hypothesis. The NH is defined by H_0 and AH is defined by H_1. Consider a basic example, we want to define whether a coin is fair or not.

Hypothesis Testing : The coin is fair or not.
H_0 : The coin is fair.
H_1 : The coin is not fair.

Now we perform experiments to check the hypothesis.

Experiment : Toss coin for 100 times.

Now if there is head-tail split is 50-50 then coin is fair, if split is 60-40 or 40-60 then also coin is fair. But if split is 30-70 or 70-30 then it is not. So coin is fair if head occurs between 40 and 60. So the [40,60] is called confidence interval. And the values beyond the CI (values < 40 and values > 60) are called significant values.

If head lies in the CI then NH is accepted otherwise it is rejected. This is called two tail test because we are checking in two directions, < 40 and > 60. CI is also called decision boundary. The CI can be defined in terms of percentage also, lets say 95%. So from 0 to 100, first 2.5% values are significant values and last 2.5% are significant values.

The SV can be standard deviation also. Lets say we want to draw conclusion for a medicine. So we can say that if a medicine's response time has standard deviation between [a,b] then NH is accepted otherwise rejected. The domain expert defines the confidence interval.

P Value :

It is the probability for the null hypothesis to be true.