2. Inferential Statistics

In Inferential statistics, we make an inference from a sample about the population. The main aim of inferential statistics is to draw some conclusions from the sample and generalise them for the population data.

E.g. we have to find the average salary of a data analyst across India. There are two options.

- The first option is to consider the data of data analysts across India and ask them their salaries and take an average.
- The second option is to take a sample of data analysts from the major IT cities in India and take their average and consider that for across India.

The first option is not possible as it is very difficult to collect all the data of data analysts across India. It is time-consuming as well as costly. So, to overcome this issue, we will look into the second option to collect a small sample of salaries of data analysts and take their average as India average. This is the inferential statistics where we make an inference from a sample about the population.

- **Population**
  In statistics, population refers to the total set of observations that can be made.
  Example: if we are studying the weight of adult women, the population is the set of weights of all the women in the world. If we are studying the grade point average (GPA) of students at Harvard, the population is the set of GPA's of all the students at Harvard.

- **Sample**
  In statistics, a sample refers to a set of observations drawn from a population.
  Often, it is necessary to use samples for research, because it is impractical to study the whole population.
  Example: We wanted to know the average height of 12-year-old American boys. We could not measure all of the 12-year-old boys in America, but we could measure a sample of boys.

- **Probability**
  Many events can't be predicted with total certainty. The best we can say is how likely they are to happen, using the idea of probability.
  Example: When a coin is tossed, there are two possible outcomes:
  1. Heads (H) or
  2. Tails (T)

  We say that the probability of the coin landing **H** is ½
  And the probability of the coin landing **T** is ½
  Formula:

$$\text{Probability of an event happening} = \frac{\text{Number of ways it can happen}}{\text{Total number of outcomes}}$$

Example: the chances of rolling a "4" with a die

Number of ways it can happen: 1 (there is only 1 face with a "4" on it)

Total number of outcomes: 6 (there are 6 faces altogether)

$$\text{So the probability} = \frac{1}{6}$$

- **Hypothesis**
  Before I get into the theoretical explanation, let us understand Hypothesis Testing by using a simple example.

  Example: Class 8th has a mean score of 40 marks out of 100. The principal of the school decided that extra classes are necessary in order to improve the performance of the class. The class scored an average of 45 marks out of 100 after taking extra classes. Can we be sure whether the increase in marks is a result of extra classes or is it just random?

  Hypothesis testing lets us identify that. It lets a sample statistic to be checked against a population statistic or statistic of another sample to study any intervention etc. Extra classes being the intervention in the above example.

  Hypothesis testing is defined in two terms – **Null Hypothesis** and **Alternate Hypothesis**.

  - **Null Hypothesis** being the sample statistic to be equal to the population statistic. For eg: The Null Hypothesis for the above example would be that the average marks after extra class are same as that before the classes.
  - **Alternate Hypothesis** for this example would be that the marks after extra class are significantly different from that before the class.

Hypothesis Testing is done on different levels of confidence and makes use of z-score to calculate the probability. So for a 95% Confidence Interval, anything above the z-threshold for 95% would reject the null hypothesis.

Points to be noted:

We cannot accept the Null hypothesis, only reject it or fail to reject it.
As a practical tip, Null hypothesis is generally kept which we want to disprove. For eg: You want to prove that students performed better after taking extra classes on their exam. The Null Hypothesis, in this case, would be that the marks obtained after the classes are same as before the classes.

Types of Errors in Hypothesis Testing:

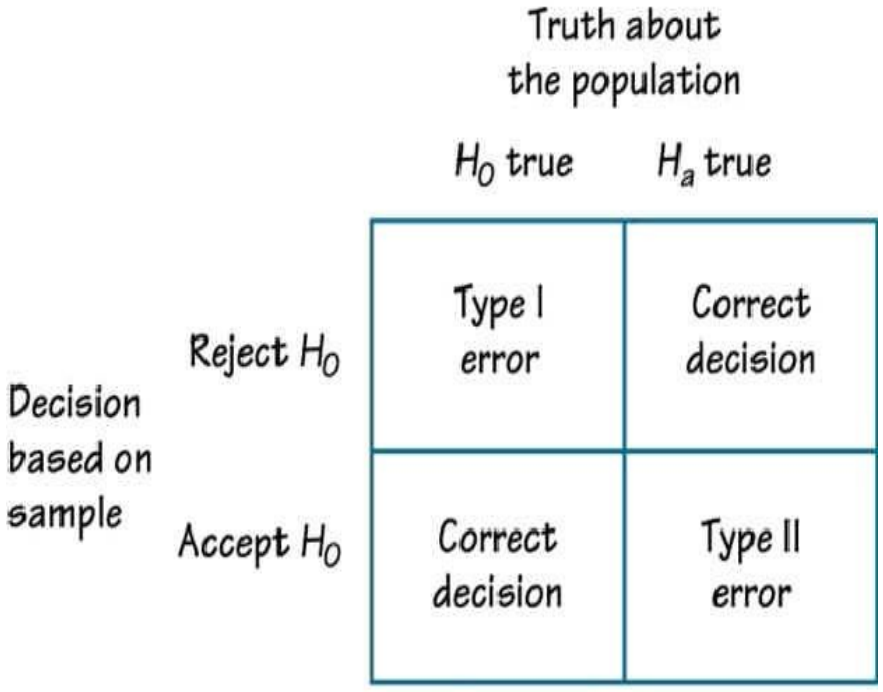
Now we have defined a basic Hypothesis Testing framework. It is important to look into some of the mistakes that are committed while performing Hypothesis Testing and try to classify those mistakes if possible.

Now, look at the Null Hypothesis definition above. What we notice at the first look is that it is a statement subjective to the tester like you and me and not a fact. That means there is a possibility that the Null Hypothesis can be true or false and we may end up committing some mistakes on the same lines.

There are two types of errors that are generally encountered while conducting Hypothesis Testing.

Type I error: Look at the following scenario – A male human tested positive for being pregnant. Is it even possible? This surely looks like a case of False Positive. More formally, it is defined as the incorrect rejection of a True Null Hypothesis. The Null Hypothesis, in this case, would be – Male Human is not pregnant.
Type II error: Look at another scenario where our Null Hypothesis is – A male human is pregnant and the test supports the Null Hypothesis. This looks like a case of False Negative. More formally it is defined as the acceptance of a false Null Hypothesis.

The below image will summarize the types of error :



- **Coefficient of Determination (R-Square)**
  It is defined as the ratio of the amount of variance explained by the regression model to the total variation in the data. It represents the strength of correlation between two variables.

  We already calculated the Regression SS and Residual SS. Total SS is the sum of Regression SS and Residual SS.

  Total SS = 2.1103632473+ 0.672210946 = 2.78257419

  Co-efficient of Determination = 2.1103632473/2.78257419 = 0.7588

- **Correlation Coefficient**
  This is another useful statistic which is used to determine the correlation between two variables. It is simply the square root of coefficient of Determination and ranges from -1 to 1 where 0 represents no correlation and 1 represents positive strong correlation while -1 represents negative strong correlation.