STATISTICS FOR DATA SCIENCE PART - 1

Introduction:

- Statistics is an applied science.
- Statistics is a branch of mathematics that deals with the modelling and analysis of data.
- In the early days Statistics is mainly used for inference.
- Later the concept of exploratory data analysis evolved in which inference is just a component.

Data:

- Data can be defined as collection of raw facts.
- The main sources of data are Images, Audio, Sensors, Click-streams…etc.
- But the major problem with this type of data is that it is unstructured.
- Unstructured data is difficult to handle by data science algorithms.
- This unstructured data must be processed and to be converted into structured data.
- In this process statistical tools help a lot.

Different types of Data:

- The different types of data are

  Continuous Data: Data that can take any values within a certain range.

  Eg: Temperature in a particular area

  Discrete Data: Data that can take only integer values.

  Eg: No of students applied for different courses

  Categorical: Data that can be categorized into different types.

  Eg: Type of a TV screen

  Binary: Data that can take only two values.

  Eg: True or False

  Ordinal: Data that has explicit ordering.

  Eg: Grades of a student

- The first two types come under numeric data whereas the last three comes under categorical data.
- Data types are useful for statistical modelling and can be thought as a signal for software like R/Python.

Rectangular Data:

- In data science data is treated in rectangular format which is similar to a table in Relational Database Management System.
- The key terminology used in rectangular data can be defined as

  Data frame: The tabular form that is used to represent the data that can be used for statistical modelling and machine learning.

  Feature: A column in the data frame can be treated as feature.

  Record: A row in the data frame can be defined as a record. Also known as tuple.

  Outcome: The output of the statistical model is defined as outcome.
- Rectangular data can be represented by using a 2 dimensional matrix.
- Both Python and R support this rectangular data concept.
- Python supports this concept with the help of pandas library which treats the rectangular data as *DataFrame* object whereas R treats this as data.frame object.
- Both provide default indexing for the rows in the data frame.
- The following pictures shows how a data frame looks like.
- This is the dataset that can be useful for heart disease prediction.
- In this the columns are age,sex,cp,…. etc which are the features of the data and we have 13 features(The last column target is ignored as it is the outcome).
- Here target is a binary variable (0 if heart disease is not present 1 otherwise).
- The task of statistical model here is to find out the relation between the first 13 columns(features) to the last column(target).
- Actually there are 303 rows though 5 are shown for convenience. Hence the size of data matrix is 303 X 13.

```
In [2]: data = pd.read_csv('heart.csv')
```

```
In [3]: data.head()
```

Out[3]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Non Rectangular Data:

- There are other data structures other than rectangular data.
- For example time series data which predicts same variable for successive intervals is not a data frame.
- Spacial data structures, graphs are also examples of non rectangular data.
- These can be modeled by using specialized models in data science.