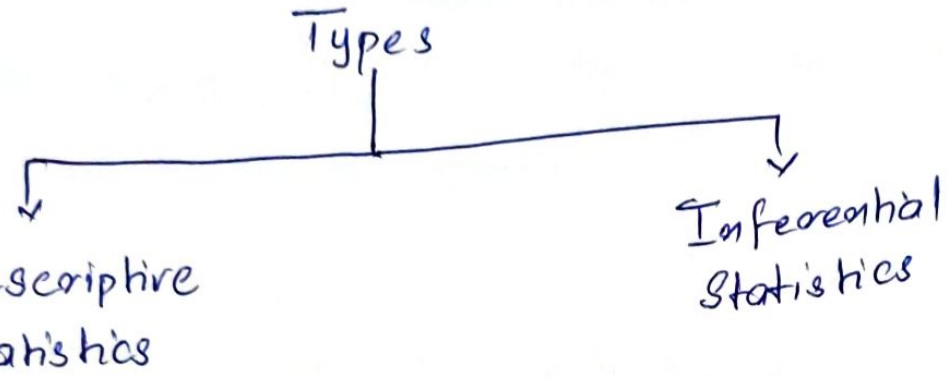
Statistics in Data Science

Statistics is a branch of mathematics, where you collect and analyze data, generally large chunks of data so that you can give some conclusions.



→ Let's take an ~~example~~ example to understand the difference b/w two types of statistics.

→ Suppose you work for HP laptop company and you are the marketing team head, and you are planning to launch a laptop which would be targeting the students in pakistan. You are planning that you will launch the laptop at a price which is attractive for a large audiance and they will immidiately buy it. But for that you should have the knowledge that in which price range students are already buying laptops. This is the challenge for you that you want to find out the average price that a student is willing to pay for his laptop in Pakistan.

way to solve this is to go to each college/university student and ask him/her how much money he/she is willing to pay for the laptop and then average it out.

→ But is it the optimal solution?

↳ Many students have distance learning education

→ This solution seems infeasible.

Alternatively,

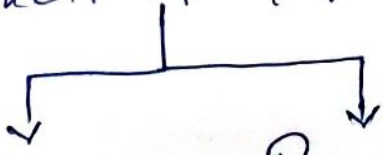
→ Instead of going to each student of each college/university, you will go to selected college students and ask them this question and compute mean of their responses, and then you will draw inferences/conclusions from this small population findings.

→ This is what is called inferential statistics and is commonly used for statistical analysis of the type we discussed.

→ Four major terms in inferential statistics

→ Population (all students of all universities)

→ Sample (segment of population we selected)



Random
(Equal chance for everyone)

Representative
(each segment of population should have representation)

ahistic (Any metric drawn on a sample space is called statistics e.g mean laptop price of sample of population

→ Parameter (Based on the mean of sample we draw conclusion about population and we call it a parameter.

→ All the metrics which are related to sample are called statistic, whereas all metrics related to population are called parameter.

→ For same metrics, formula is different for population and sample. For example mean of population is denoted by μ and mean of sample is denoted by $\bar{x}$. Same is case for standard deviation, variance. and many others.

→ Sample is denoted by n whereas population is denoted by N.

Descriptive Statistics:

→ Simplest kind of statistics

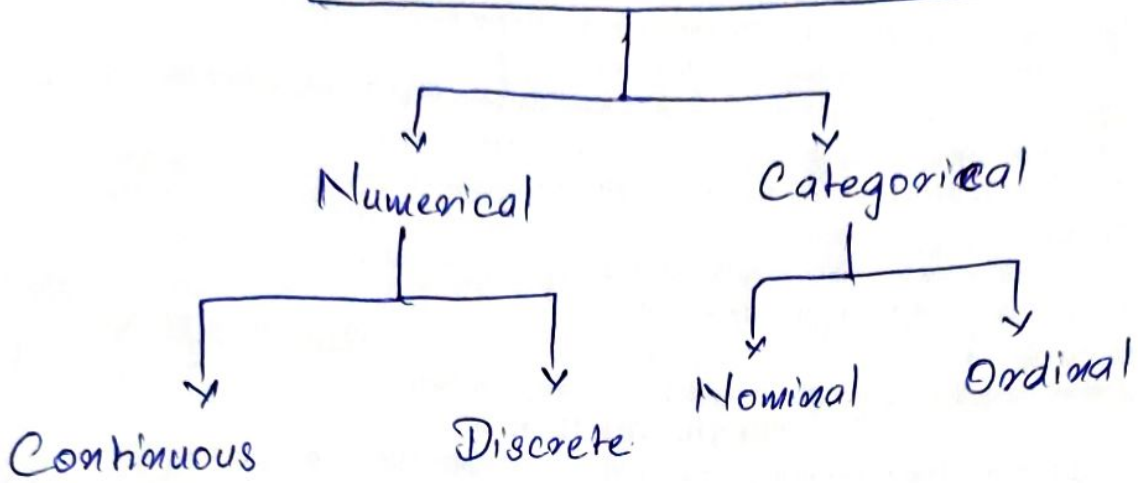
→ Also known as summary statistics.

- Mean
- Median
- Mode
- Standard deviation
- Variance
- Co-variance.

of the topics included in inferential statistics ④

→ Confidence Intervals
→ Hypothesis Testing.

Types of Data in Statistics



→ Price of smartphone is a type of numerical data.

→ Brand of smartphone is a type of categorical data.

→ Price of smartphone is a type of numerical continuous data. Numerical continuous data can achieve any value in a given range. e.g temperature.

→ Number of applications installed on your smartphone is numerical discrete data. Numerical discrete data can have a specific set of values.

Nominal Categorical Data is one in which there is no order b/w the categories. For examples gender. No one can say Male is greater than Female or Female is greater than Male.

Ordinal Categorical Data is one in which there is an order b/w the categories. For example if I say how you will rate your mobile phone battery performance among bad, average and best. and we know average is better than bad so there is an order b/w categories.

Studying different types of data

→ There are two types of data

- → Numerical
  - → Continuous
  - → Discrete
- → Categorical
  - → Nominal
  - → Ordinal

egorical Data:

For example we have geander data

Geander
M
F
M
F
O
M
M
M
F

→ We draw freq table

Geander Type	Count
M	23
F	57
O	3

↓

To analyze this type of data we can use.

↳ Bar Chart

↳ Pie Chart

Numerical Data:

Suppose we have age data.

24, 14, 31, 47, ----- 90

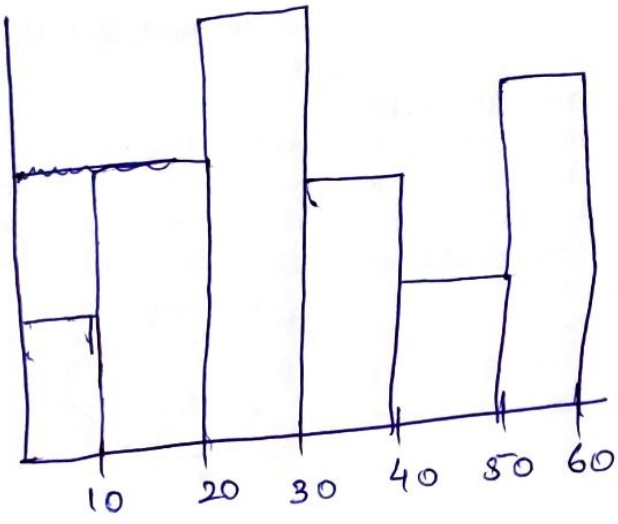
→ To visualize this's type of data is little different.

→ A bit of different technique is used to draw freq table.

Instead of counting each value seperately, we create bins. like 1-10, 11-20, 21-30 to draw freq table

Freq Table:

Type	Count
1-10	4
11-20	7
21-30	12.



→ It is called histogram

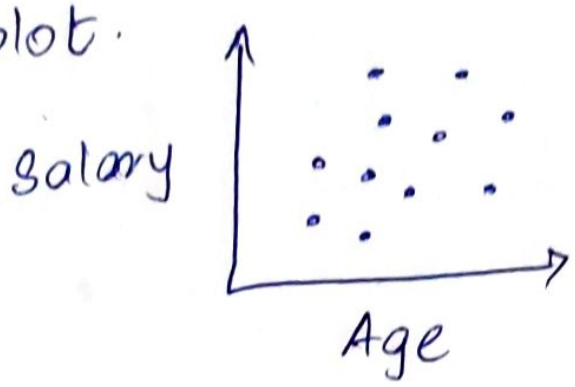
→ It is used when we have continuous numerical data, where category is bins.

→ When we have discrete numerical data it can be represented by a bar chart.

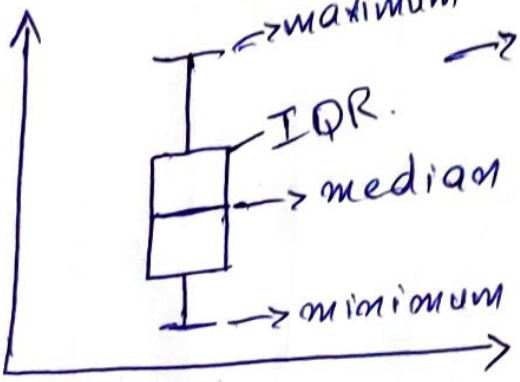
→ Univariate Analysis (Analysing one column of data at one time)

→ Multivariate / Bivariate / Trivariate Analysis (Analysing more than one columns at a time, it can be 2, 3, or more)

What if you have two different columns of continuous numerical data you use scatter plot.



→ If you want to see the variation of continuous data you can use boxplot.



→ Used for univariate analysis.