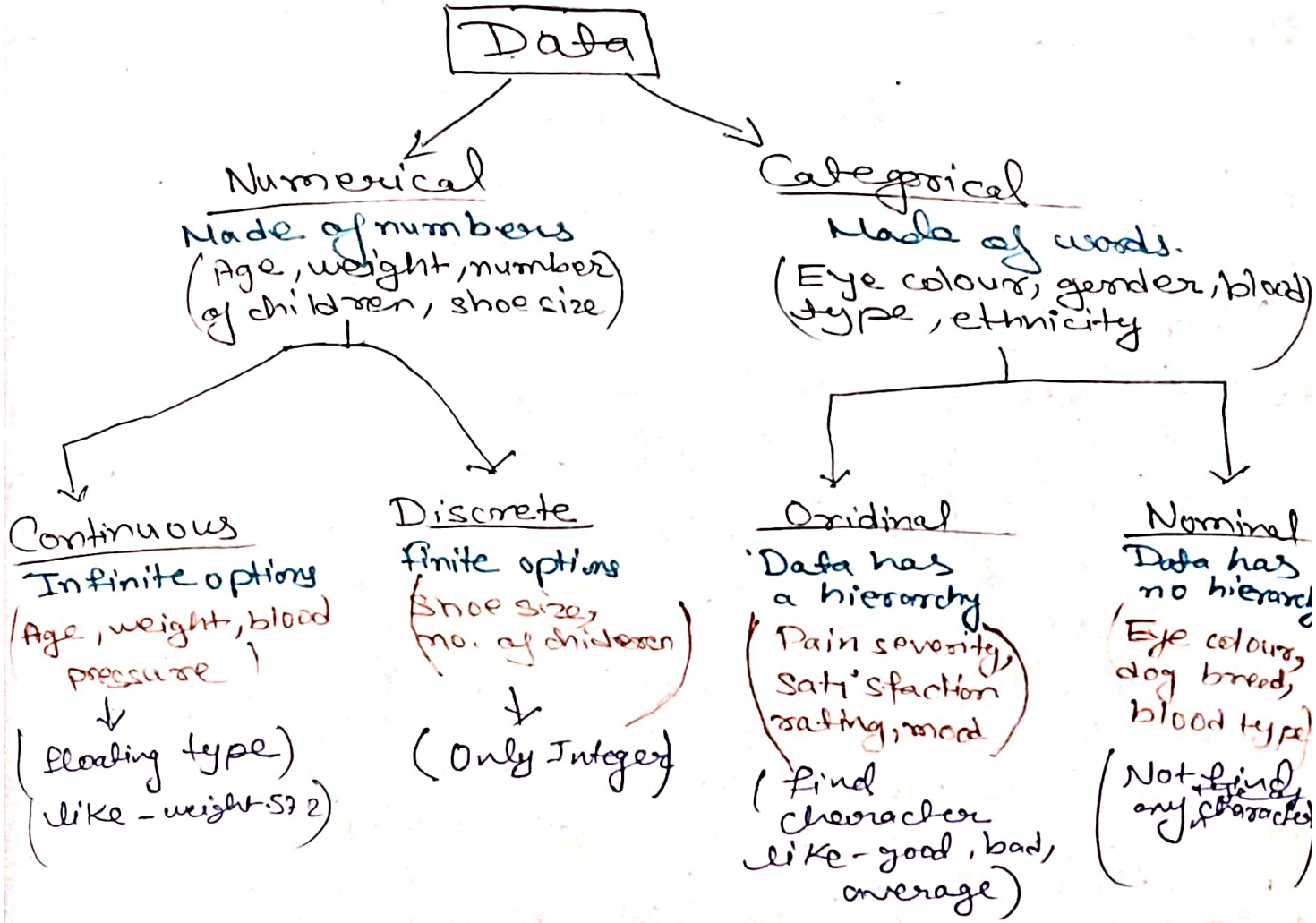
STATISTICS

Data → Data are the facts and figures collected, summarized, analyzed, and interpreted. The data collected in a particular study are referred to as the data set.



① Categorical Data → Categorical data represents characteristics. Therefore it can represent things like a person's gender, language etc. It can also take on numerical values. ~~Eg:-~~

② Nominal Data → Nominal ~~Data~~ Values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as "labels". Note that nominal data that has no order. Therefore, if you would change the order of its values, the meaning would not change the order of its values, the meaning would not change. You can see two examples of nominal features below:

1) Are you married?
○ Yes ○ No

2) What languages do you speak?
○ English ○ Chinese
○ French ○ Spanish

(b) Ordinal Data – Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters. You can see an ex —

What is Your Educational Background?
○ 1 ⇒ Elementary ○ 2 ⇒ High School
○ 3 ⇒ Undergraduate ○ 4 ⇒ Graduate

Numerical Data

(a) Discrete Data – We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data can't be measured but it can be counted. It basically represents information that can be categorized into a classification. An example is the no. of heads in 100 coin flips

(b) Continuous Data

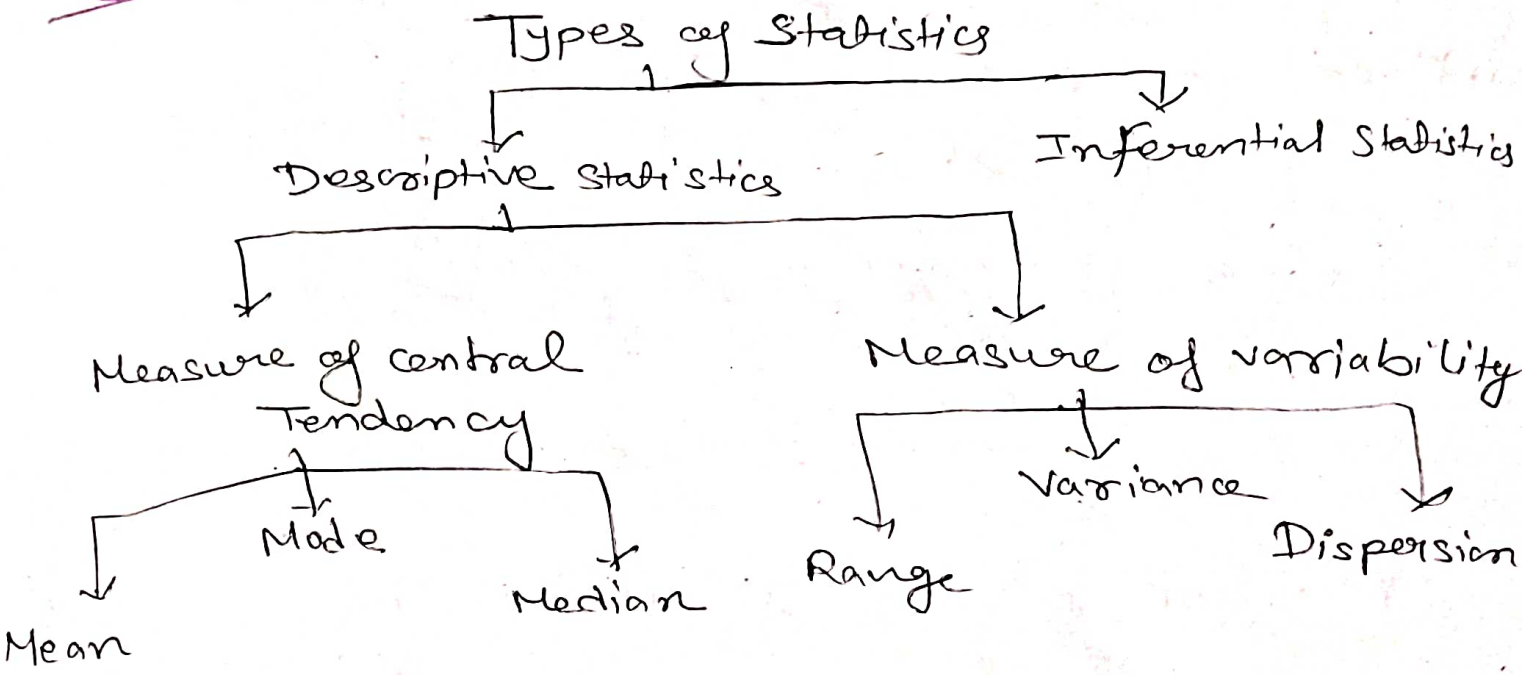
⊙ Continuous Data represents measurements and therefore their values can't be counted but they can be measured.

Continuous data is further divided into two categories: —

(i) Interval data – Interval data type refers to data that can be measured only along a scale at equal distance from each other. eg ⇒ body temperature can be measured in degree celsius and degree Fahrenheit and neither of them can be 0.

(iii) Ratio data — unlike interval data, ratio data has zero point. Being similar to interval data, zero points is the only ~~different~~ difference they have. eg → in the body temperature, the zero point temperature can be measured in Kelvin.

Types of Statistics



Descriptive Statistics → In the descriptive Statistics the data is described in summarized way. ~~The summarization is done from~~ It uses data that provides numerical calculation or graph or table.

Inferential Statistics ⇒ It makes inference and prediction about population based on a sample of data taken from population. It generalizes a large dataset and applies probabilities to draw a conclusion. It is simply used for explaining meaning of descriptive ~~stat~~ stats. It is simply used to analyze, interpret result, and draw conclusions.

Measure of central tendency

It is also known as summary statistics that is used to represents the center point or a particular value of a data set or sample etc. In statistics, there are three common measures of central tedency

(i) **Mean** It is measure of average of all value in a sample set.

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5

$$\text{Mean}(m) = \frac{\text{Sum of all the terms}}{\text{Total no. of terms.}}$$

$$= \frac{21.3 + 20.8 + 19}{3} = 20.366$$

$$\text{Mean} = \frac{\text{Sum of values of each observation}}{\text{Number of observation.}}$$

In ~~pyth~~ numpy module:

```
import numpy as np
list1 = np.random.randint(3,10,20)
np.          (list1)
```

Output
5.35

list1

array([4, 6, 4, 7, 6, 5, 3, 3, 8, 8, 6, 9, 3, 8, 9, 9, 8, 9, 5, 5])

How to find mean value in csv file ?

```
sd = pd.red_csv('stock-data.csv')
```

sd.close

Output
40.91
40.97
.
.
.
3018 69.85

np.mean(sd.close) → Output 28.41274260

In all column value is array

(iii) Median → It is measure of central value of a sample set. In these, data set is ordered from lowest to highest value and then finds exact middle.

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5
i20	15	4

Ordering the set from lowest to highes = 15 19 20.8 21.3

$$\text{Median} = \frac{19 + 20.8}{2} = 25.5$$

Median is the middle value of the distribution when the values are arranged in ascending or descending order.

In numpy:

```
list1 = np.random.randint(3,10,20)
np.median(list1)
```

How do you find median in any file.

```
nd = pd.excel('newdata.xlsx')
np.median(nd.work-exp)
3.8499....
```

(iii.) Mode → It is value most frequently arrived in sample set. The value repeated most of time in central set is actually mode.

2 3 4 2 4 6 4 7 7 4 2 4

Mod = 4

In numpy there is no direct function to find mode. So we will use statistics module to use it.

```
import numpy as np
import statistics as st
list = np.random.randint(3,10,20)
st.mode(list)
```

Mean()

~~n = [5,5,3,5,2,7]~~

~~if n>0:~~

~~l = len(n)~~

~~if (l>0):~~

~~r = n[0],~~

2, 2, 5, 5, 5, 7

~~27/6 = 8~~

~~5+10/2 = 8~~

8

Mean = 8 | Mode = 5

Median = 3

5, 5, 3, 5, 2, 7, 29 (* outlayer)

Mean → 8 | 2, 3, 5, 5, 5, 7, (29)

Median → 5 | Mode → 5

Mean is not stable in central ~~t~~endency because it is more effected to ~~previous one~~ change data.

When outlayer is present in given data then null value can't be change by mean. You can change median and mode.