Statistics

|  |  |
| --- | --- |
| Numerical | Categorical |
| Quantitative | Qualitative |
| Discrete and Conti. |  |

Levels of Data

1. Nominal level ====== > Categorical
2. Ordinal level ======= > Categorical
3. Interval level ======= > Numerical
4. Ratio level ======= > Numerical

Temp is the best example you cant take the ratio, because tem does not have zero scale

Conclusion:

Interval level no zero scale

Ratio level zero scale

Population vs sample

Will work on sample ===== > Population ==== > Inferential statistics

What are table names and graphical names

How do you represent the data i.e categorical data and Numerical data

Categorical data : English statement

Ex: In a class there 50 members girls and 30 members boys are there

You need to represent as a table format

|  |  |
| --- | --- |
| Boys | Girls |
| 30 | 50 |

|  |  |
| --- | --- |
| Gender | Total |
| Boys | 30 |
| Girls | 50 |

Groups are here class

Boys are one class Girls are another class

Total we represent as frequency

Frequency: Number of cycles per second

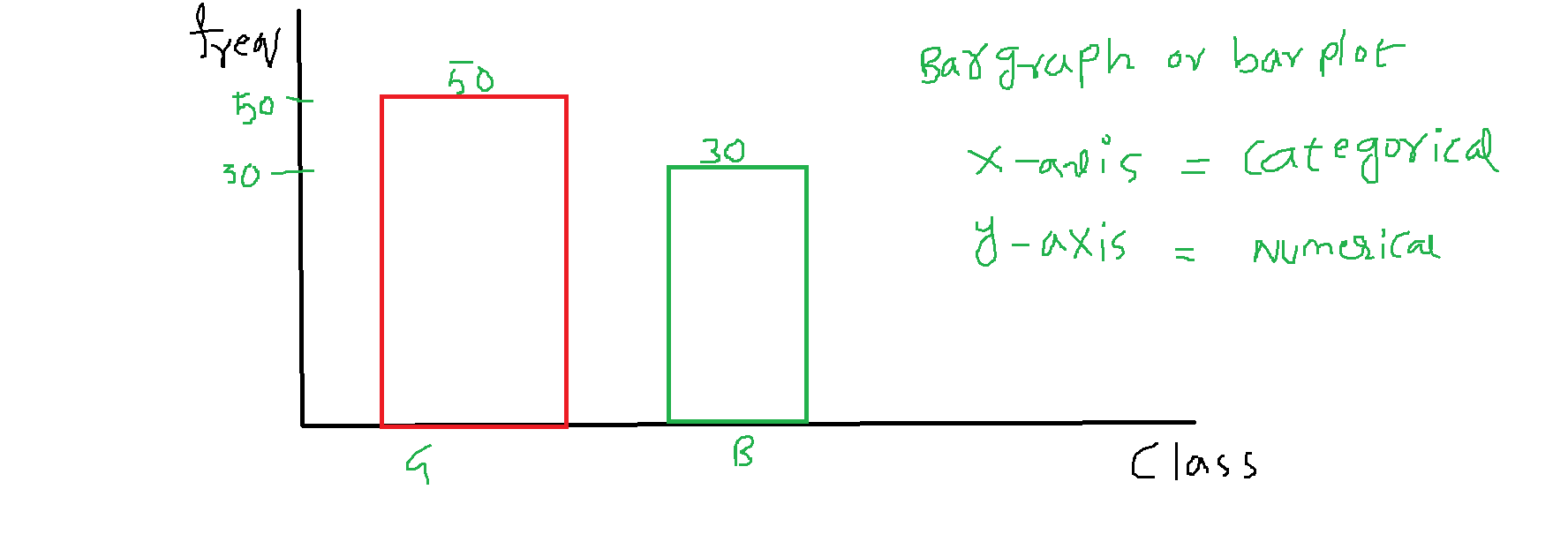
**Frequency table:**

|  |  |
| --- | --- |
| Class | Frequency |
| Boys | 30 |
| Girls | 50 |

If you take any data set

|  |  |
| --- | --- |
| Class | Frequency |
| Asia | 30 |
| Africa | 50 |
| Europe |  |

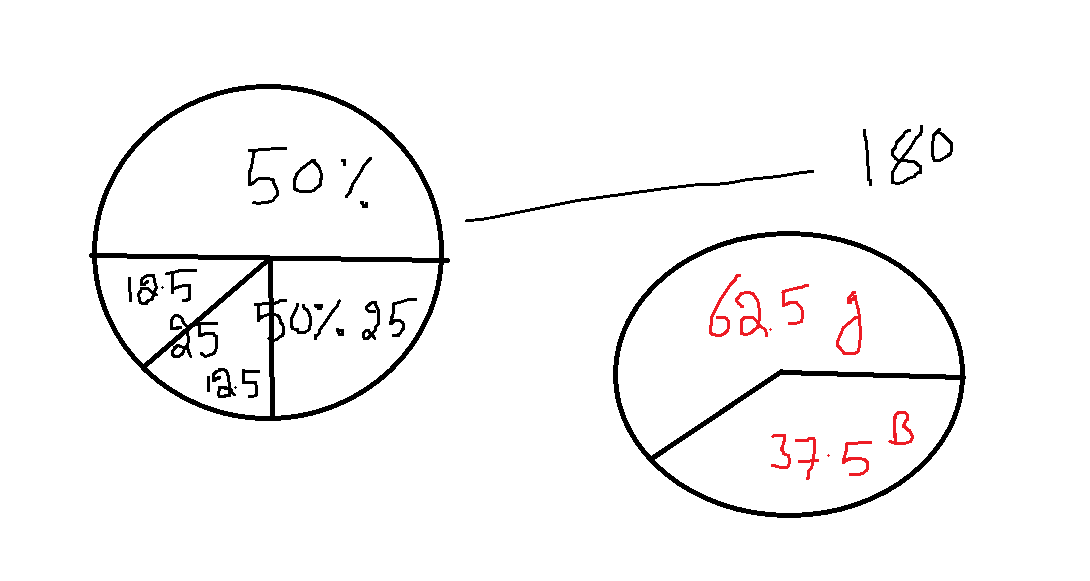
**Graphical representation of Frequency table: Bar chart**



**Relative Frequency table:**

|  |  |  |
| --- | --- | --- |
| Class | Frequency | Relative Frequency |
| Boys | 30 | (30/80)\*100=37.5% |
| Girls | 50 | (50/80)\*100=62.5% |
|  | 80 | 100% |

**Graphical Representation of Relative frequency table : Pie-Chart**



Summary:

Categorical data tabular format name: Frequency table

Categorical data graphical representation: Bar chart

X-axis: Class y-axis: Frequency

Categorical vs Numerical

Categorical data greaphical representation relative frequency table: Pie chart

|  |  |
| --- | --- |
| Tabular | Graphical |
| Frequency table | Bar chart |
| Relative frequency table | Pie chart |

80% EDA 20% ML

Numerical data :

Suppose in a class we have 10 members are there in that 6 members Boys and 4 Girls are there

Marks of 10 members students: Raw data

5 8 8 10 12

15 15 17 19 20

We divide the data into some groups, will count how many person are there in that count

So that we can know

**Frequency distribution table:**

|  |  |
| --- | --- |
| Class interval | Interval frequency |
| 0-5 | 1 |
| 5-10 | 3 |
| 10-15 | 3 |
| 15-20 | 3 |

How to divide the data into parts

How many parts you need to divide

How much gap of interval

0-3 0

3-6 1

6-9 2

9-12 1

12-15 3

15-18 1

18-20 2

0-2 information 0-3

More intervals is a problem also less intervals is a problem

20k> above : 5

25k> above: 2

30k> above: 1

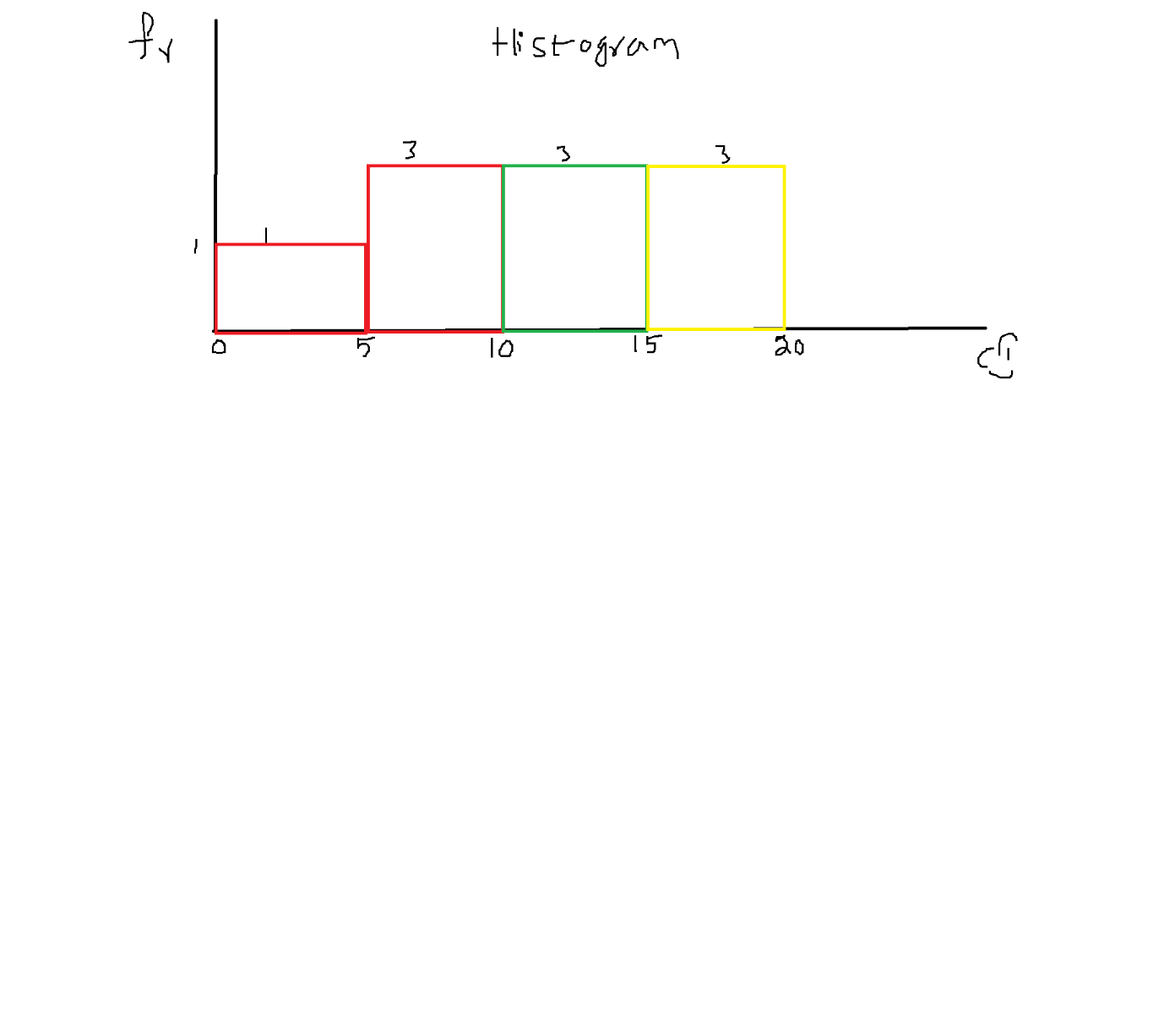
We have some rules are there

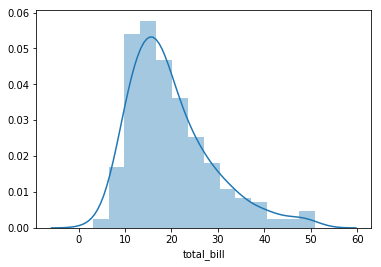
Graphical representation of Frequency distribution table: **Histogram**

**X-axis: class interval (numerical column)**

**y-axis: interval frequency (numerical column)**

**numerical vs Numerical**

****



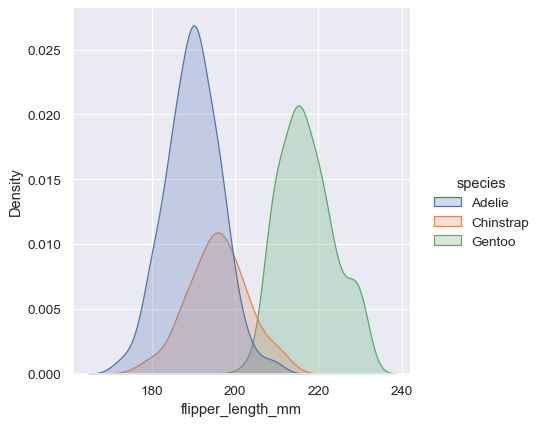
Distribution plot is enhanced from histogram

Distribution gives more clear idea about how data is distributed i.e. data flow

How data is flow / distributed / dispersed from one point to another point

To make a distribution plot

* First we need plot a histogram
* To achieve the histogram we need to identify class interval and its frequency
* To achieve this , we need to divide raw to into suitable intervals



Comparing distribution plots easy than comparing histograms

Raw data ===== > CI ===== > Interval feq ==== > Freq distribution table ==== > Histogram ===> Distribution plot

Numerical data:

|  |  |
| --- | --- |
| Table | Graph |
| Frequency distribution table | Histogram  X-axis: Numerical CI  Y-axis: Numerical frequency |
|  | Distribution plot |