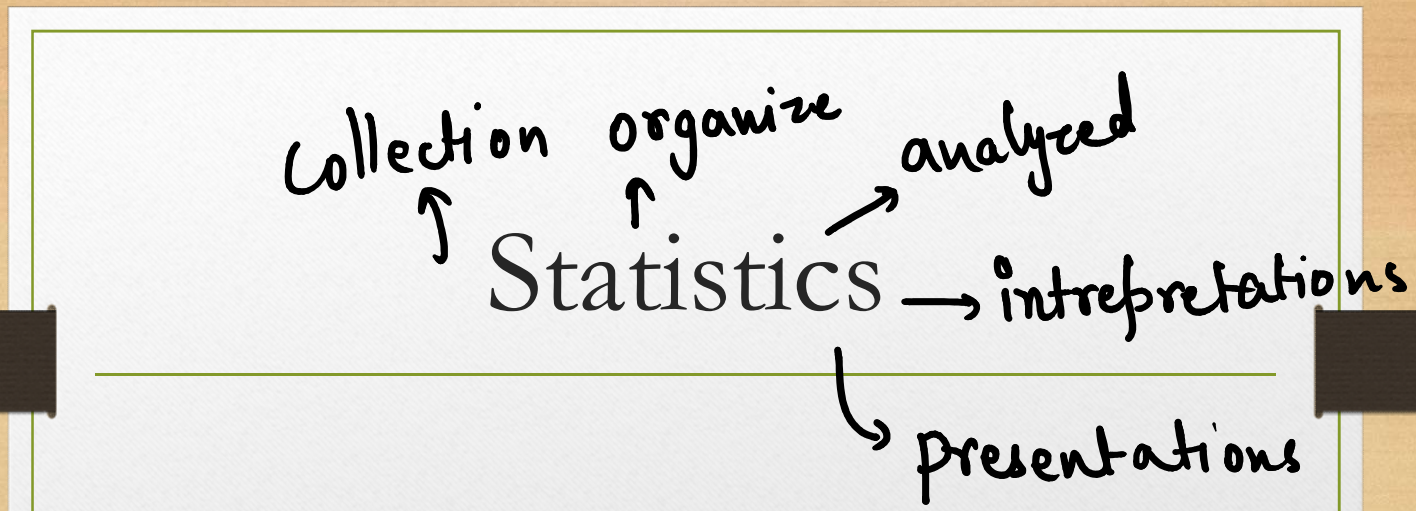
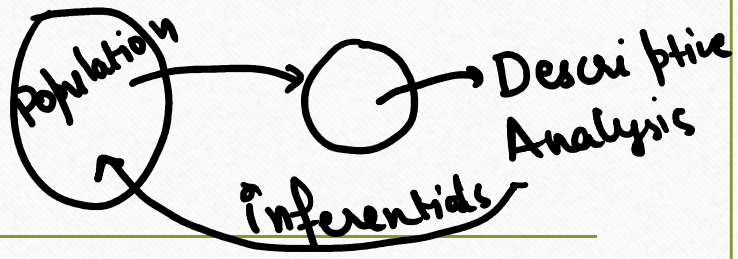
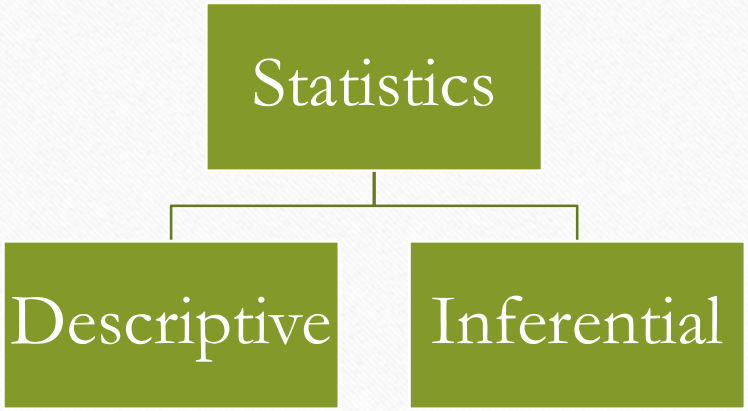
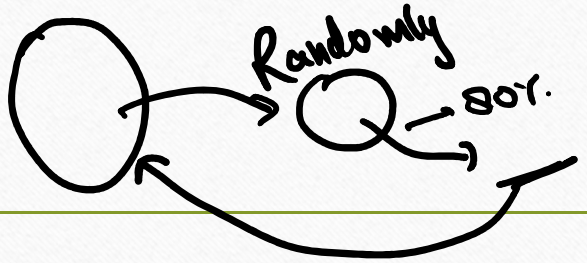

collection
organize
analyzed
Statistics
intrepretations
presentations

Statistics

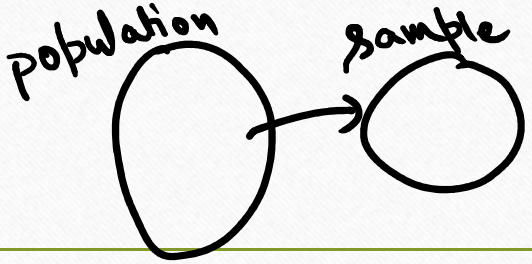


- Statistics are numbers that summarize raw facts and figures in some meaningful way. They present key ideas that may not be immediately apparent by just looking at the raw data and facts or figures from which we can draw conclusions. Statistics can be a convenient way of summarizing key truths about data.





Types

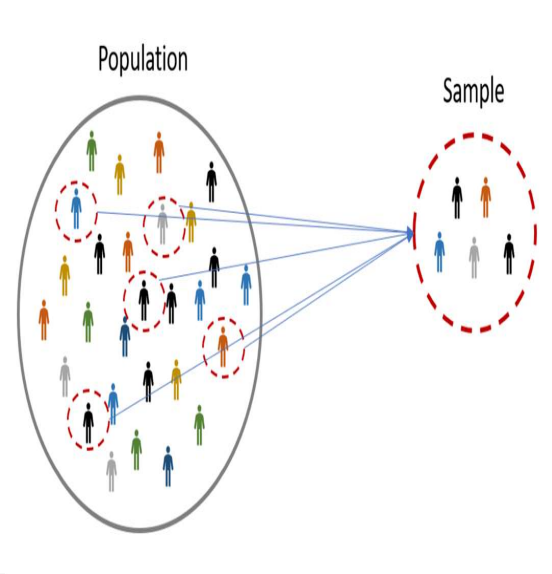


- Descriptive Statistics: Descriptive statistics describe, show, and summarize the basic features of a dataset found in a given study, presented in a summary that describes the data sample and its measurements. It helps analysts to understand the data better. We don't draw any inferences from data.

- Inferential Statistics: Inferential statistics takes a random sample of data from a portion of the population and describes and makes inferences about the entire population.

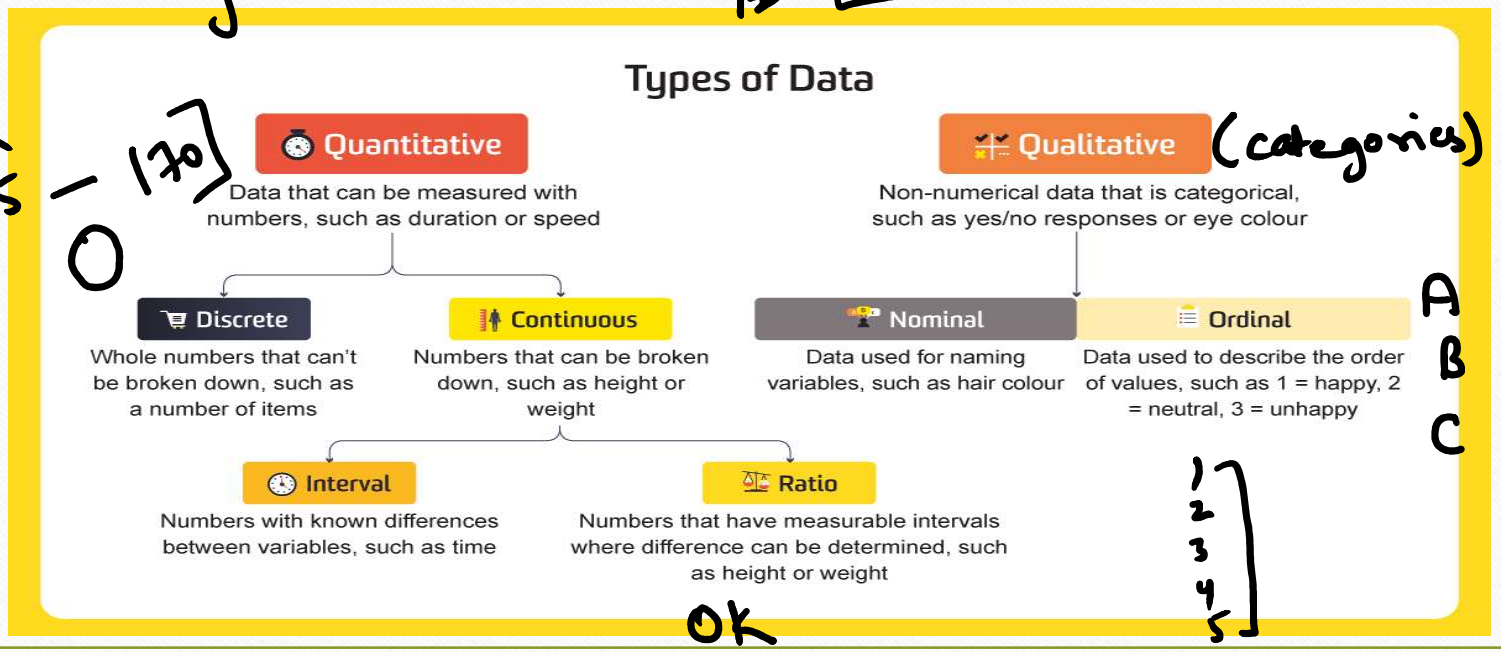
Basic Concepts

- Population: Total number of data points or things.
- Sample: Small subset of population used for study.
- Sample Size: Numbers of data points present in sample.
- Variable: Any feature or attribute of object under study.



Variables

1 year 3 months ⇒ $1 + \frac{3}{12} = 1.25$



(categories)

A B C

1 2 3 4 5

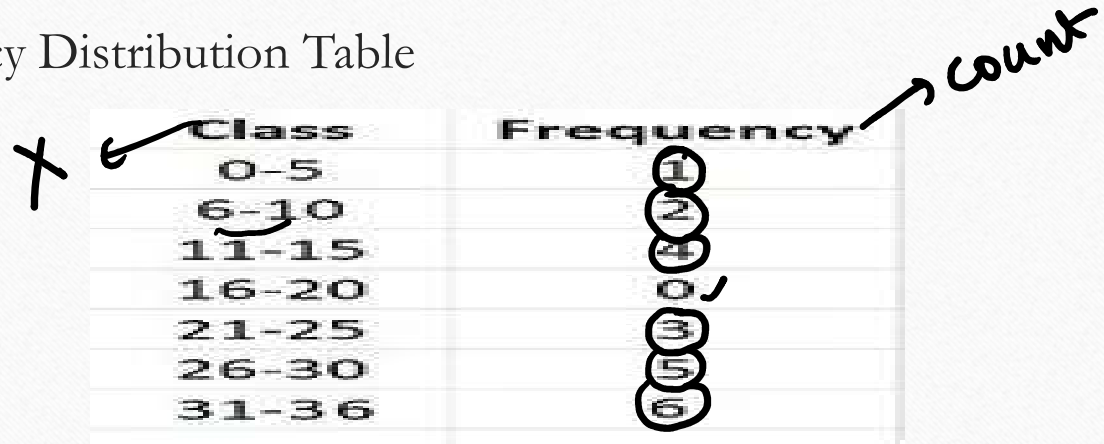
OK

Descriptive Statistics

- Frequency Distribution
- Measure of Central Tendency: Mean, Median and Mode
- Measure of Spread: Range, Variance, Standard deviation and IQR
- Measures of Symmetricity: Kurtosis and Skewness
- Five Number Summary

Frequency Distribution

- Frequency: It tells us **how often something happened**. The frequency of an observation tells you the number of times the observation occurs in the data.
- Frequency Distribution Table



Class	Frequency
0-5	1
6-10	2
11-15	4
16-20	0
21-25	3
26-30	5
31-36	6

How to create Frequency Table?

- Step 1: Make a table with two columns—one with the title of the data you are organizing, like grades, scores, temperatures, etc., and the other column for frequency.
- Step 2: Look at the items written in the data and decide whether you want to draw an ungrouped frequency distribution table or a grouped frequency distribution table. If there are many different values, we should go with the grouped frequency distribution table.
- Step 3: Write the data set values in the first column.
- Step 4: Count the number of times each item is repeating itself in the collected data. In other words, we have to find the frequency of each item by counting.
- Step 5: Write the frequency in the second/third column corresponding to each item.
- Step 6: Write the total frequency in the last row of the table

FDT for Ungrouped Data

- Below are the scores of 35 students in a science test (out of 10). Arrange these in a tabular form using tally marks: [5, 8, 7, 6, 10, 8, 2, 4, 6, 3, 7, 5, 8, 5, 1, 7, 4, 6, 3, 5, 2, 8, 4, 2, 6, 4, 2, 8, 9, 5, 4, 7, 5, 5, 8.]

Scores in Science	Tally Marks	Frequency
1	I	1
2	IIII	4
3	II	2
4	𝍸	5
5	𝍸 II	7
6	IIII	4
7	IIII	4
8	𝍸 I	6
9	I	1
10	I	1
Total		35

(No. of students)

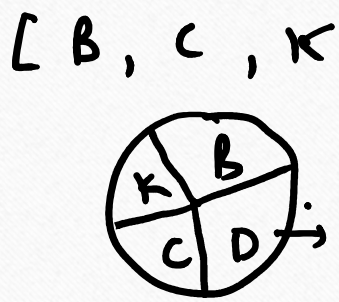
Frequency distribution Table

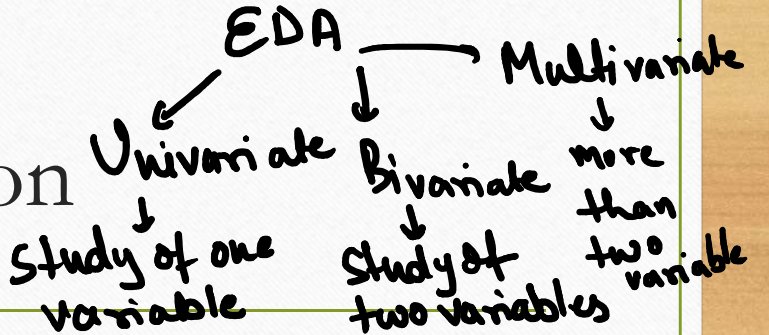
FDT for Grouped Data

- The following are age groups of 20 people in a concert: 5, 65, 62, 48, 5, 23, 17, 40, 32, 34, 35, 51, 6, 18, 52, 28, 39, 41, 20, 69. Construct a grouped distribution table with class intervals 0–10, 10–20 and so on.

Class Interval	Frequency
0-10	3
10-20	2
20-30	3
30-40	4
40-50	3
50-60	2
60-70	3
Total	20

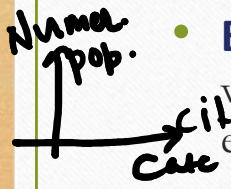
Data Visualization





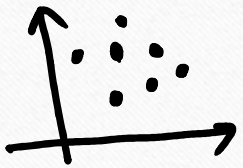
- **Line Chart -** Line charts are a fundamental chart type generally used to show change in values across time.

- **Pie Chart -** Pie charts are used to compare the proportions of different groups or categories.
- **Bar Graph -** each bar represents a particular category, and the length of the bar indicates the value. The longer the bar, the greater the value. All the bars have the same width, which makes it easier to compare them.



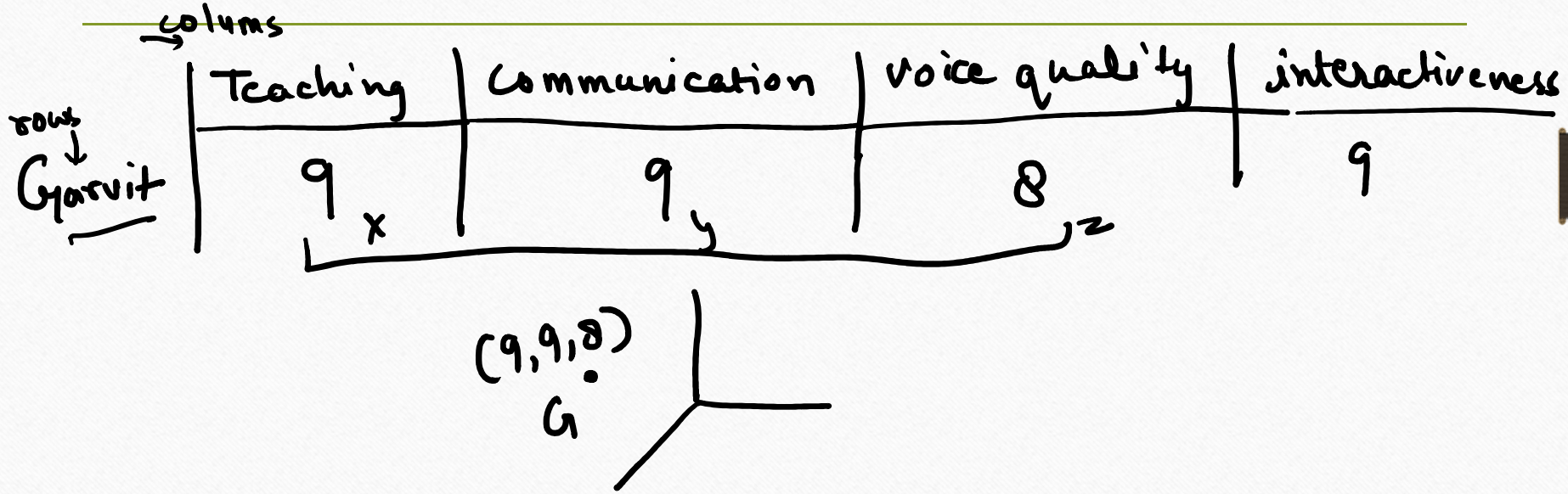
Bar has higher precision than pie.

- **Histogram -** used to plot frequency distributions. → univariate
- **Scatter Plot -** used to plot two different numerical features.



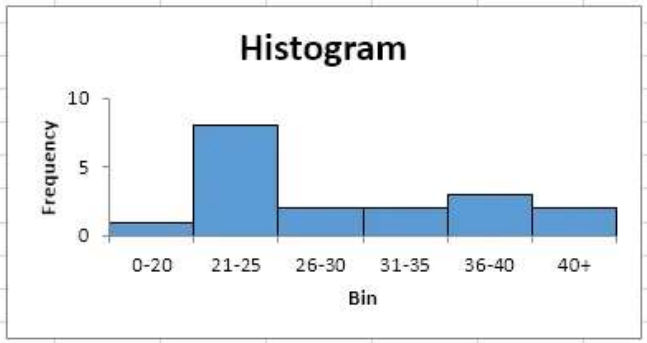
Columns → variables, features, dimension, fields etc.

rows → datapoints, record, entries etc.



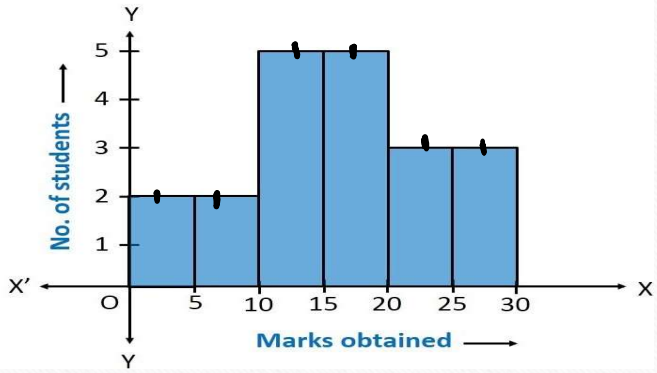
Building Histogram from FDT

Bin	Frequency
0-20	1
21-25	8
26-30	2
31-35	2
36-40	3
40+	2



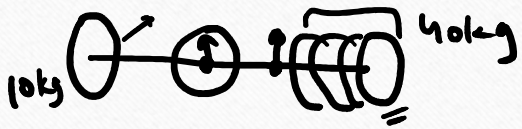
Building Histogram from FDT

Intervals	Frequency
0 - 5	2
5 - 10	2
10 - 15	5
15 - 20	5
20 - 25	3
25 - 30	3



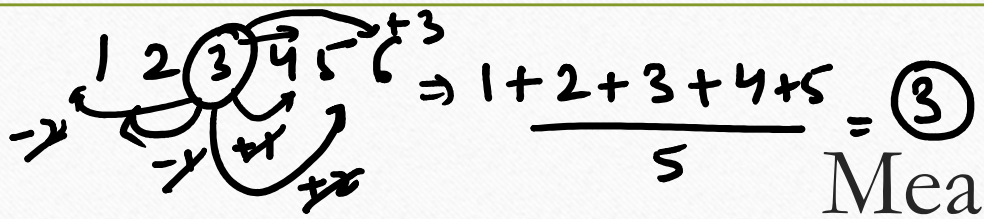
Outliers

- Abnormal observation or data point present in the data. It is an extreme value.
- It can be exceptionally high or exceptionally low.
- Always glance through the data to keep check on it.
- They are present in the data because of unintentional errors or natural errors.



Measures of Central Tendency

- Average
  - Mean
  - Median
  - Mode



$\Rightarrow \frac{1+2+3+4+5}{5} = 3$

Mean 10 = [10 10 10 10]

Mean

np.mean() [10, 10, 10, 10, 100]

= Mean $\Rightarrow \frac{140}{5} = 28$

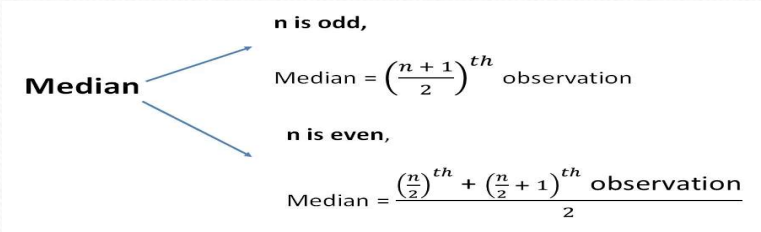
- The most common type of Average out there.
- Most influenced by the outliers.
- It changes when symmetricity of the distribution/graph is affected.

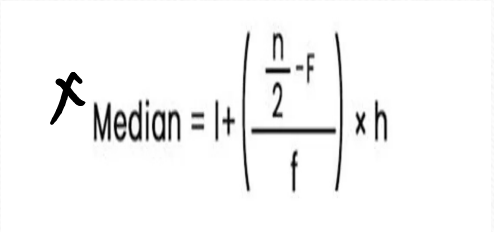
$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\bar{x} = \frac{\sum f x}{\sum f} \Rightarrow \text{frequency distribution table}$$

Median

- It is actually the middle value, dividing the data into 2 equal halves.
- Steps:
  - Sort the data in ascending order





- It is not influenced by outliers.

$\Rightarrow [10, 10, 10, 10] \Rightarrow$ Median = 10

$\frac{\frac{4^{th}}{2} + \left(\frac{4}{2}+1\right)}{2} \Rightarrow \frac{2^{th} + 3^{th}}{2} = \frac{10+10}{2} = 10$

$[10, 10, 10, 10, 100] \Rightarrow$ Median = 10

$[10, 10, 10, 10, 100, 200, 300, 400] \Rightarrow$ Median = $\frac{10+100}{2} = 55$

Missing values (numerical column)

if your column have outliers, then use median.

if your column has no " , then use mean.

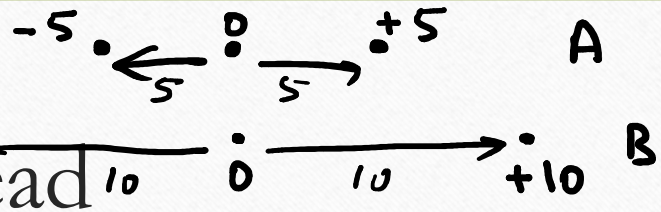
Mode

- It works for both numerical and categorical data.
- Observation with highest frequency is the mode.
- A dataset can be unimodal or multimodal.
- For ungrouped data, we can count or direct observe in the table.
- For grouped data, we use

$$\text{Mode} = l_1 + \left(\frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \right) \times c$$

Measures of Spread

$\text{Mean}_A = 0$
$\text{Median}_A = 0$
$\text{Mean}_B = 0$
$\text{median}_B = 0$



- Measures
  - Range
  - Variance
  - Standard Deviation
  - Quartile Range(IQR)

$R_A = 5 - (-5) = 10$

$R_B = 10 - (-10) = 20$

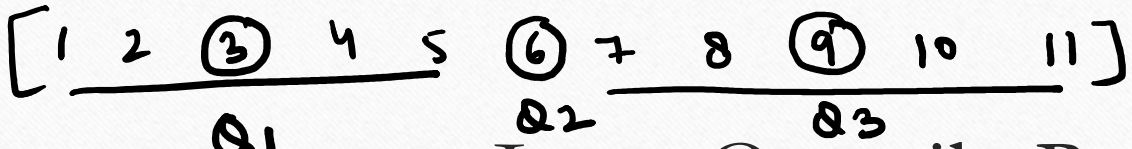
Range

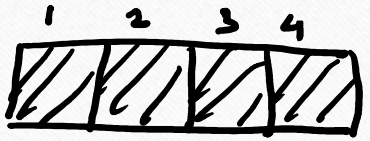
$[-5, 0, 5, 500]$

$R = 500 - (-5) = 505$

- The range is a way of measuring how spread out a set of values are.
- The range only describes the width of the data, not how it's dispersed between the bounds.
- Range is very sensitive to outliers.

$$\text{Range} = X_{max} - X_{min}$$





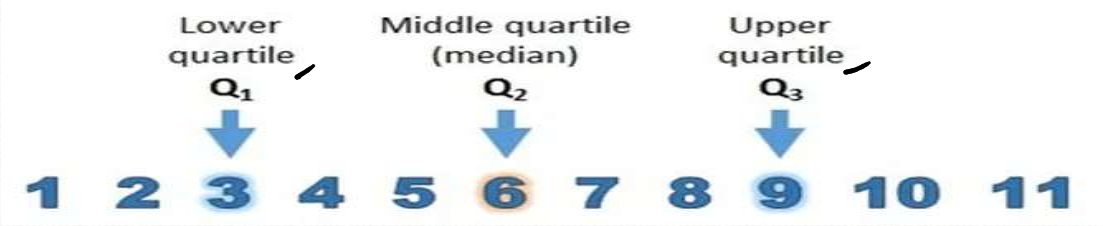
Q1 Q2 Q3

25% 50% 75%

Inter Quartile Range

$IQR = Q_3 - Q_1 = 9 - 3 = 6$

- Quartiles: The numbers that separate data into 4 equal parts.



- IQR: difference between the 3rd quartile and 1st quartile.
- Less sensitive to outliers

$$IQR = Q_3 - Q_1$$

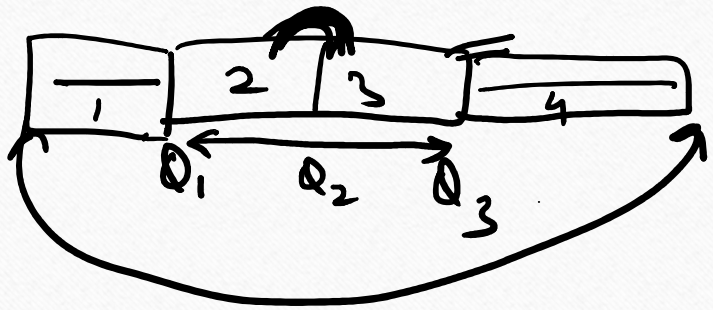
[1, 2, ③, ④, 5, ⑥, ⑦, 8, 9, 10, 11, 12]

6.5

$Q_2 = \frac{6+7}{2} = 6.5$

$Q_1 = \frac{3+4}{2} = 3.5$

$Q_3 = \frac{9+10}{2} = 9.5$



$IQR = Q_3 - Q_1 = 9.5 - 3.5 = 6$

Q ≡ [17, 25, 17, 20, 19, 18, 23, 64, 33, 22]

↓ sort

21 (Q_2)

[17, 17, (18) 19, 20, 22, 23, (25) 33, 64]

1st — find

Sol. ≡

$Q_2 = 21$

$Q_1 = 18$

$Q_3 = 25$

$IQR = 25 - 18 = 7$