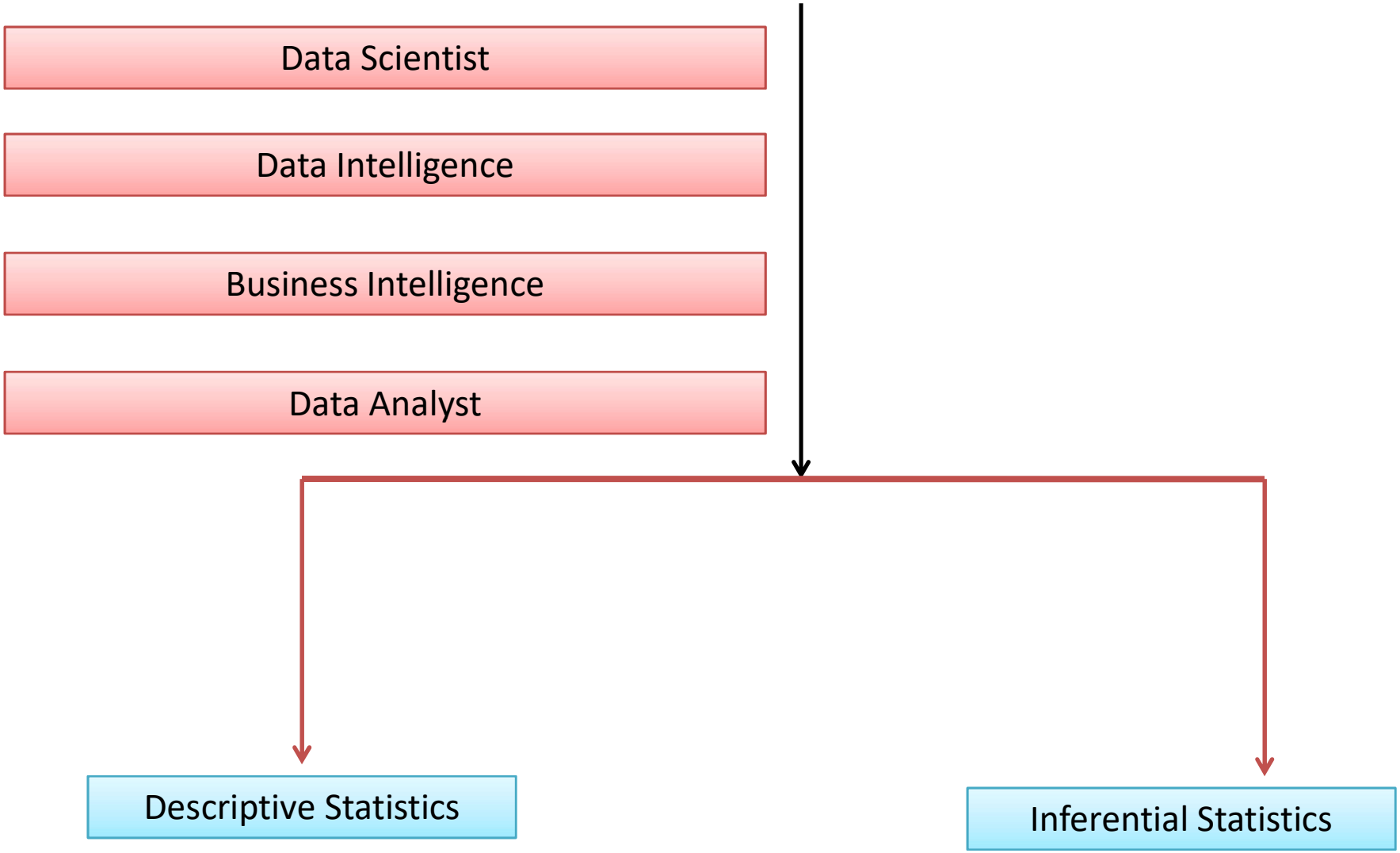
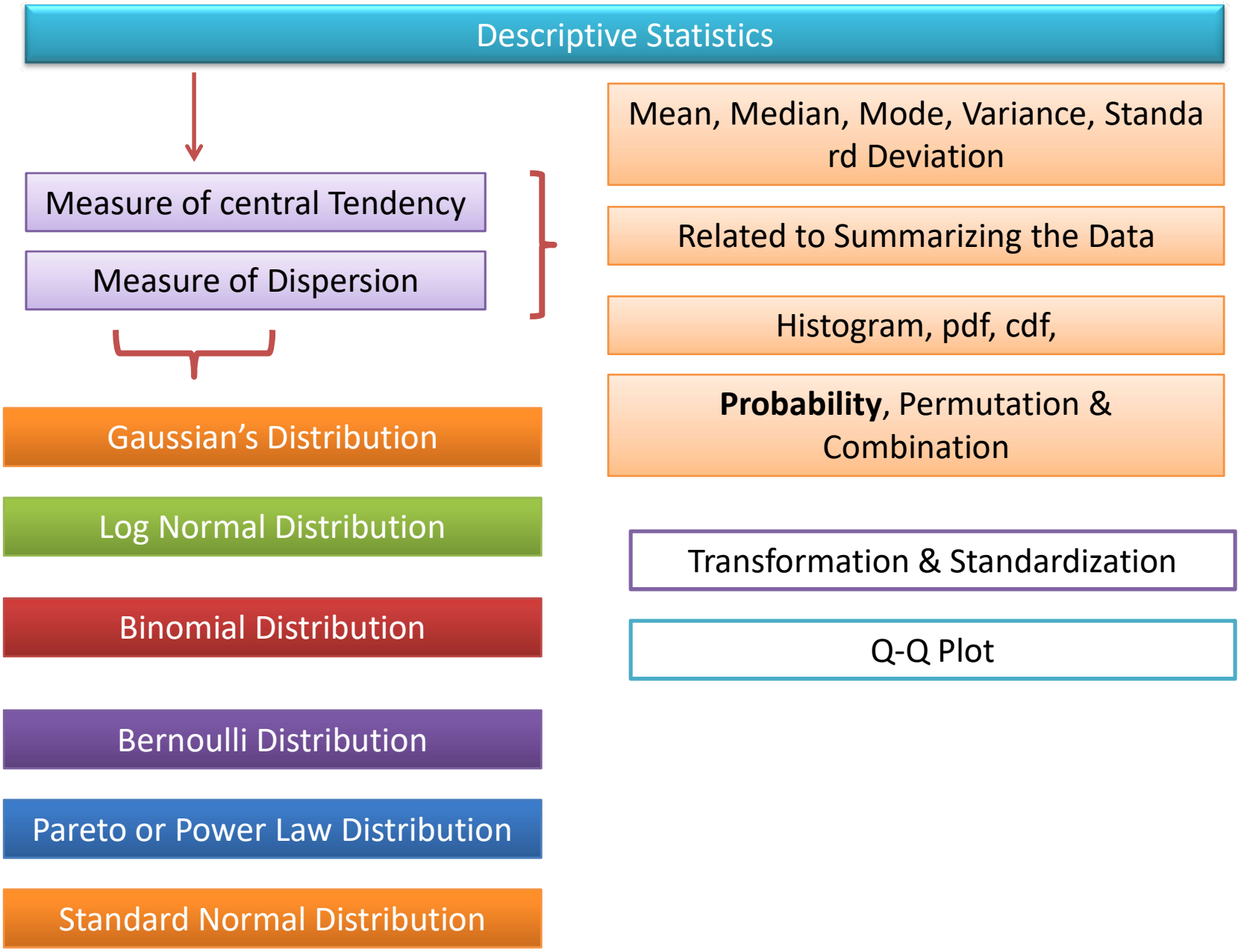
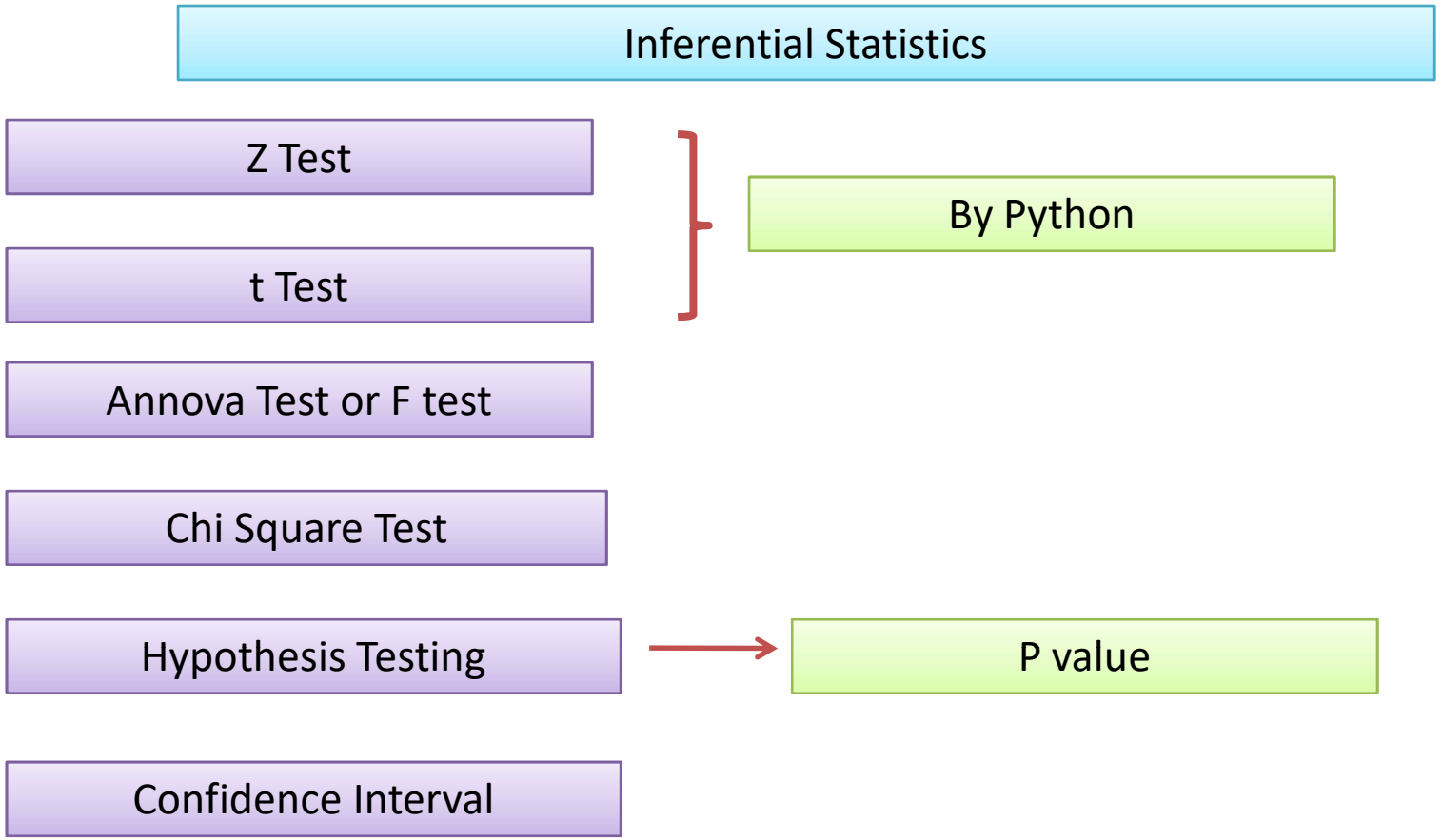
COMPLETE STATISTICS FOR DATA SCIENCE




Descriptive Statistics
Measure of central Tendency
Measure of Dispersion
Mean, Median, Mode, Variance, Standa rd Deviation
Related to Summarizing the Data
Histogram, pdf, cdf,
Probability, Permutation & Combination
Gaussian's Distribution
Log Normal Distribution
Binomial Distribution
Bernoulli Distribution
Pareto or Power Law Distribution
Standard Normal Distribution
Transformation & Standardization
Q-Q Plot


Inferential Statistics
Z Test
t Test
By Python
Annova Test or F test
Chi Square Test
Hypothesis Testing
P value
Confidence Interval

What is Statistics ?

Science of Collection, organizing and analyzing Data. → Used for **BETTER DECESSION MAKING**

What is Data?

Facts or Process of Information that can be measured. → Ex: Ages of Student of Class **30, 25, 20,28,20 (Data)**

Types of Statistics ?

1. Descriptive Stats → It consists of organizing and summarizing of Data

2. Influential Stats → Technique where in we use the **DATA**, that we have measured to form **CONCLUSION**

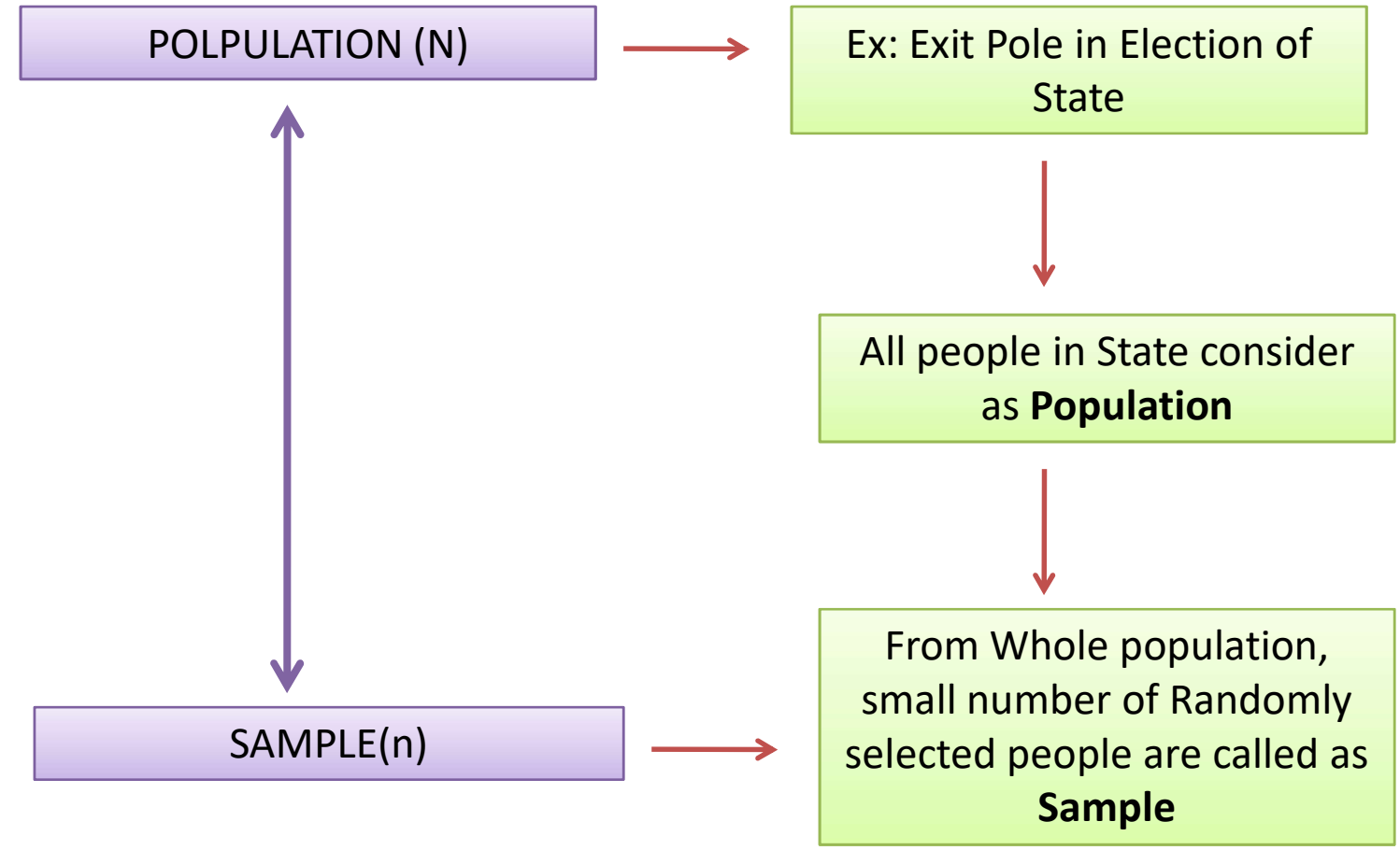
Example of Descriptive stats

Class room of Math Students: 20 No.
Marks of Ist Semester
84,85,89,90,75,55, 84,85,89,90,75,55 84,85,89,90,75,55,35,40
What is the Average Marks of Students of class room?

Example of Influential stats

Are the Average marks of the students of Maths classroom similar to the marks of Math Class in college?

POPULATION & SAMPLE



METHOD OF SAMPLING TECHNIQUES

1. Simple Random → Mostly used; Ex. Exit Pole opinion of Election

2. Stratified Sampling → Population (N) is split into **Non Overlapping Groups (Strata)**
Example:
a. Gender (Male/ Female)
b. Survey Based on Specific age group

3. Systematic Sampling → Pick up every nth (specific number) individual from population

4. Convenient Sampling →
- Survey from specific domains
- Related to specific Topic

What is Variable?

Variable is the property that can have different value.
Ex: Height, Age, Body Weight

Two Type

Quantitative

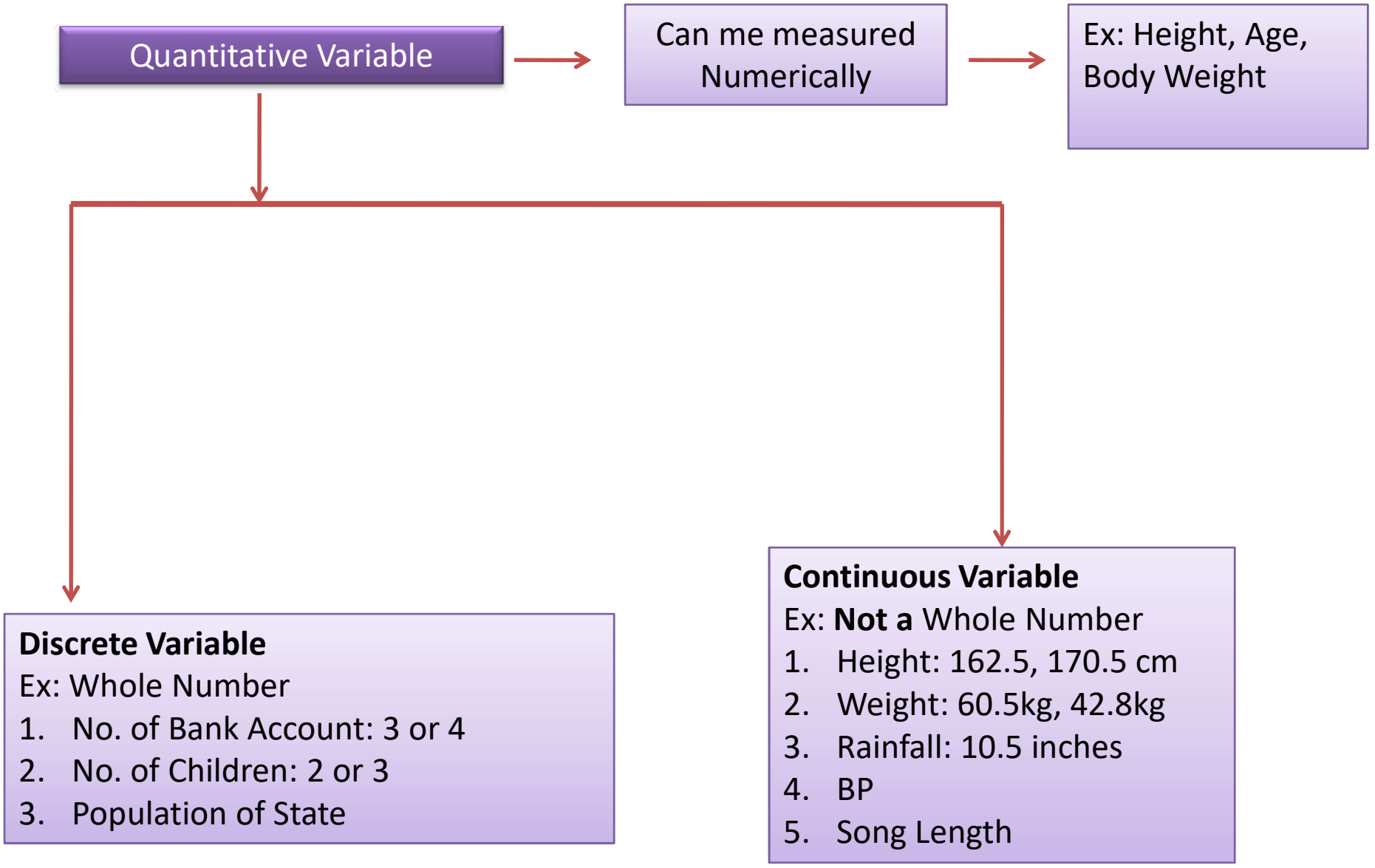
Can me measured Numerically

Ex: Height, Age, Body Weight

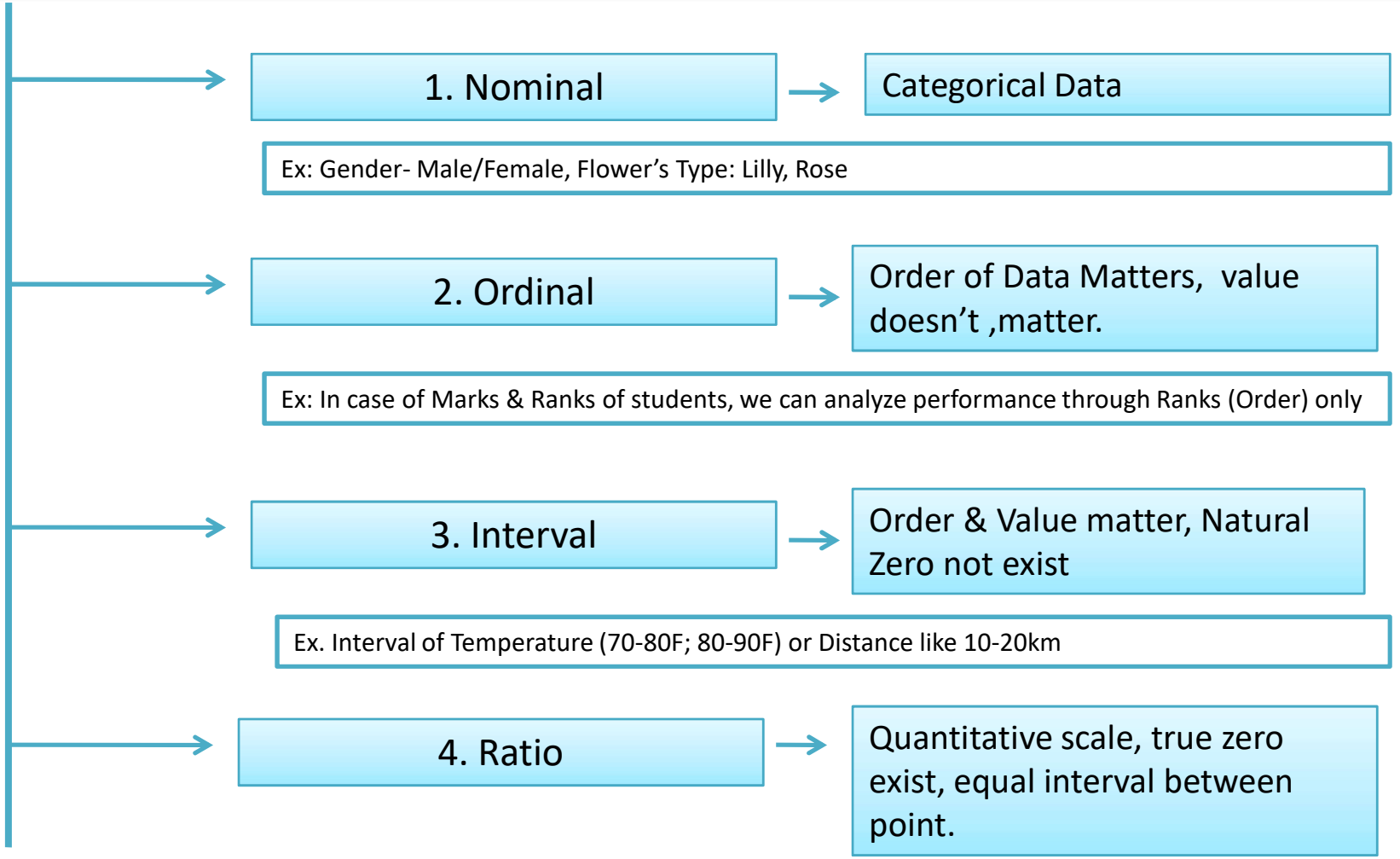
Qualitative/ Categorical

Based on characteristic, we can derived categorical variable

Ex: Gender (Male/Female)


Quantitative Variable
Can me measured Numerically
Ex: Height, Age, Body Weight
Discrete Variable
Ex: Whole Number
1. No. of Bank Account: 3 or 4
2. No. of Children: 2 or 3
3. Population of State
Continuous Variable
Ex: Not a Whole Number
1. Height: 162.5, 170.5 cm
2. Weight: 60.5kg, 42.8kg
3. Rainfall: 10.5 inches
4. BP
5. Song Length

Variable Measurement Scale

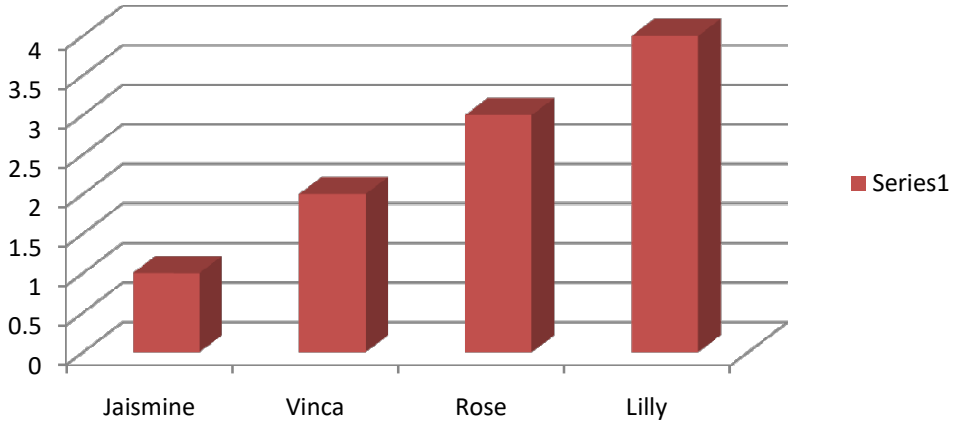


Frequency Distribution

Ex: Flowers- Rose, Lilly, Rose, Rose, Vinca, Jaismine, Vinca, Lilly, Lilly, Lilly

Flowers	Frequency	Cumulative Frequency
Rose	3	3
Vinca	2	5
Lilly	4	9
Jaismine	1	10

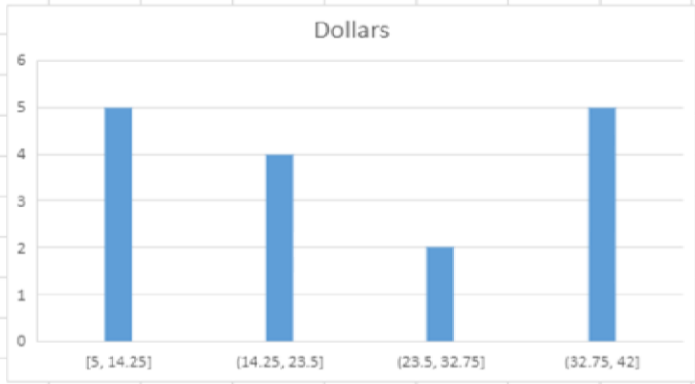
Bar Chart



Value is Discrete Variables

Histogram: Data Should be continuous Variable

Name of Student	Dollars
MISRI	5
KATHA	6
SHIVANI	8
ABHISHEK	15.5
KARTIK	18
NEHA	12
ISHAN	11.6
IVAN	19
HIBA	30
MEGHA	35
TAVISH	33.5
SHLOKA	38
AKANSHA	40
PREM	42
SANKET	21
DHARA	25

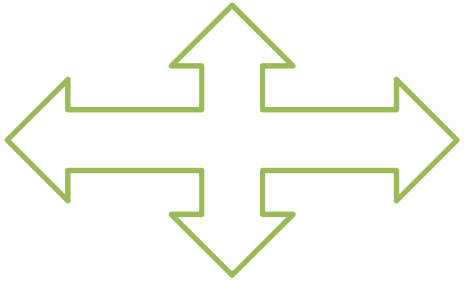


Note:
In above Histogram, we have selected **4** bins

MEASURE OF CENTRAL TENDENCY

Refer to the measure used to determine the central of distribution of data

Arithmetic **MEAN** for Population(N) and Sample(n)



Arithmetic **MODE** for Population(N) and Sample(n)

Arithmetic **MEDIAN** for Population(N) and Sample(n)

→ Median work well with **OUTLIER**

Arithmetic **MEAN** for Population(N) = Calculate Average only

Arithmetic **MEDIAN** for Population(N) and Sample(n)= Select Central or middle element of data

Arithmetic **MODE** for Population(N) and Sample(n)= Select most frequent Element of data

MEASURE OF CENTRAL TENDENCY

Example of Data: MODE

Suppose we have data with some missing elements

Type of Flower	Petal Length
Rose	
Vinca	
Lilly	
Rose	
Vinca	
Rose	
?	
?	

Missing elements: can be determined with modes

↓

Arithmetic **MODE** for Population(N) and Sample(n)= Select most frequent Element of data

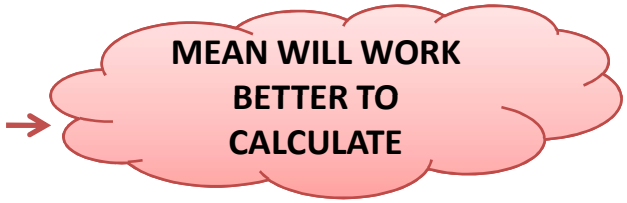
Mode will work with Categorical Variable

MEASURE OF CENTRAL TENDENCY

Example of MEAN

Suppose we have data (Quantitative variable) with some missing elements

Name of Student	Age
MISRI	30
KATHA	15
SHIVANI	35
ABHISHEK	35
KARTIK	?
NEHA	30
ISHAN	?



MEASURE OF DISPERSION [SPREAD]

- VARIANCE → Spread of Data
- STANDARD DEVIATION → Square Root of Variance

EXAMPLE OF TWO DATASET

DATASET: 1	1	1	2	2	4
DATASET: 2	2	2	2	2	2

MEAN OF BOTH DATASET IS SAME that is **5**

So, How we can distinguish Distribution ????

It can be distinguish with the help of VARIANCE

PERCENTILE & QUARTILE

Use to find Outliers or Odd numbers in Data

PERCENTILE is the value, below which a certain percentage of observation lie.

Ex: Data Set

Q1 . = What is the Percentile Ranking of 10?

2	2	3	4	5	5	5	6	7	8	8	8	8	8	9	9	10	11	11	12

Percentile value of 10 = $\frac{\text{No. of value below 10}}{\text{n (sample size)}} \times 100$

Percentile value of 10 = $\frac{16}{20} \times 100 = 80\%$

PERCENTILE & QUARTILE

Ex: Data Set

Q 2 . = What value exist at Percentile Ranking of 25%?
(Reverse of Question 1)

2	2	3	4	5	5	5	6	7	8	8	8	8	8	9	9	10	11	11	12

Value = $\frac{\text{Percentile}}{100} \times (n+1)$

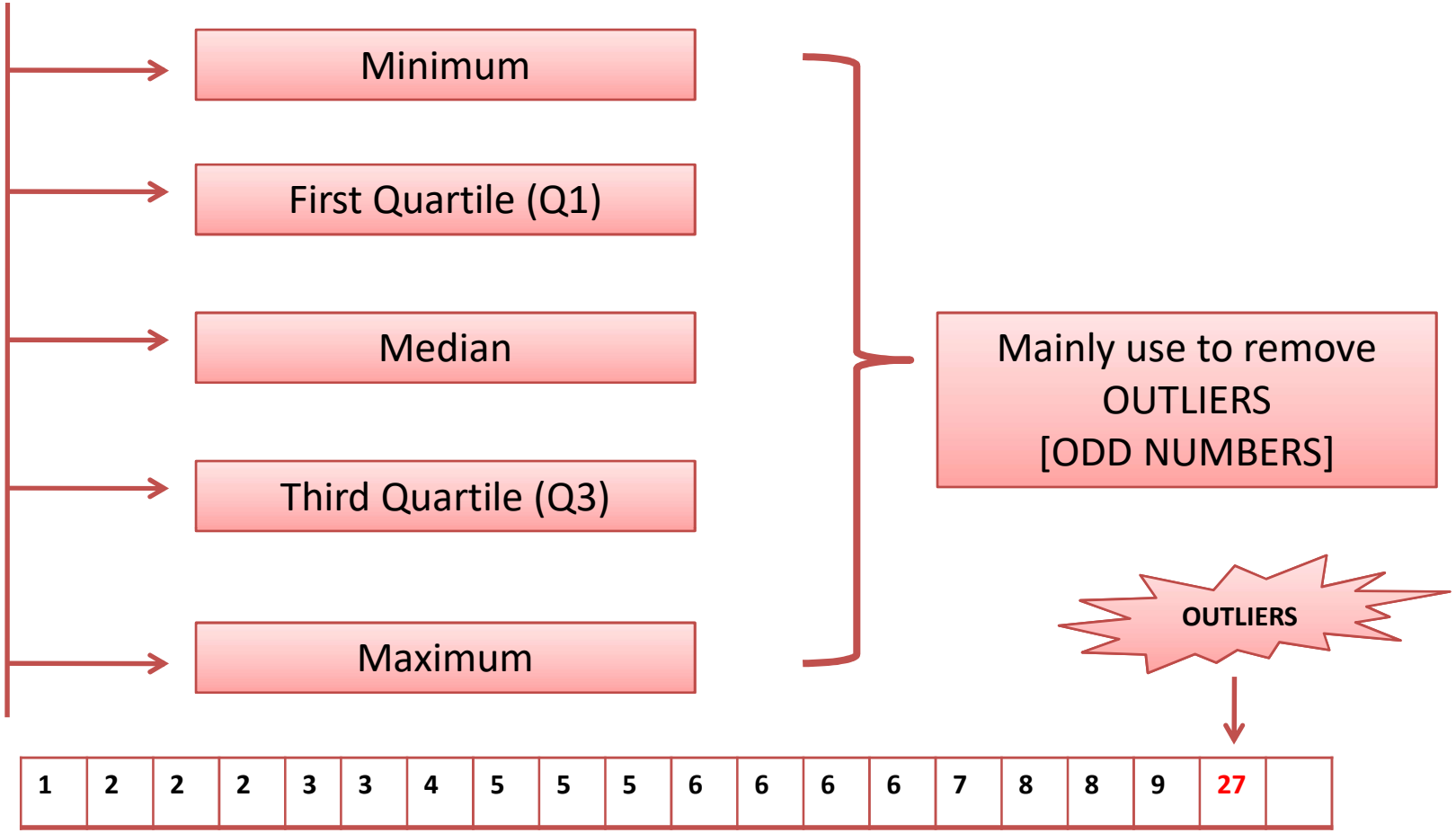
Solution: Value =
$\frac{25}{100} \times (20+1) = 5.25$ (This is INDEX value)

2	2	3	4	5	5	5	6	7	8	8	8	8	8	9	9	10	11	11	12
1	2	2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Index value → Index value 5.25 comes between 5 & 6 index → Will take Average of Index value of 5 & 6

In this case: Index value of 5 & 6 is 5,5: Average of 5+5/2=5

FIVE NUMBER SUMMARY



FIVE NUMBER SUMMARY

How to Remove Outliers from following Data Set?

1	2	2	2	3	3	4	5	5	5	6	6	6	6	7	8	8	9	27	

Answer: We need to define LOWER FENCE & HIGHER FENCE

Lower Fence: Below the lower fence all number will be treated as Outliers

Lower Fence= Q1-1.5 (IQR)

Upper Fence: Above the Upper/Higher fence all number will be treated as Outliers

Upper Fence= Q3+1.5 (IQR)

Where:
IQR: Inter Quartile Range
IQR: Q3-Q1
Q1: 25 Percentile
Q3: 75 Percentile

FIVE NUMBER SUMMARY: REMOVAL OF OUTLIERS

Solution:

Data Set

1	2	2	2	3	3	4	5	5	5	6	6	6	6	7	8	8	9	27	

1	2	2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Index Number

Q1=25 Percentile

$\frac{25}{100} * (n+1)$

$\frac{25}{100} * (19+1)$

Result:5 or Index No. 5

Data in index 5 is “3”

Q3=75 Percentile

$\frac{75}{100} * (n+1)$

$\frac{75}{100} * (19+1)$

Result:15 or Index No. 15

Data in index 15 is “7”

Upper Fence= Q3+1.5 (IQR)

Lower Fence= Q1-1.5 (IQR)

IQR= Q3-Q1

IQR= 7-3=4

FIVE NUMBER SUMMARY: REMOVAL OF OUTLIERS

Solution: Continue…….

Upper Fence= Q3+1.5 (IQR) → Upper Fence= 7+1.5 (4) → Upper Fence= 13

Lower Fence= Q1-1.5 (IQR) → Lower Fence= 3 -1.5 (4) → Lower Fence= -3

Data Set

1	2	2	2	3	3	4	5	5	5	6	6	6	6	7	8	8	9	27	

As per definition of OUTLIER, Numbers that exist below the lower fence and those exist above Upper fence in Data Set are consider as OUTLIERs.

OUTLIERS

Data Set after removal of OUTLIER

1	2	2	2	3	3	4	5	5	5	6	6	6	6	7	8	8	9

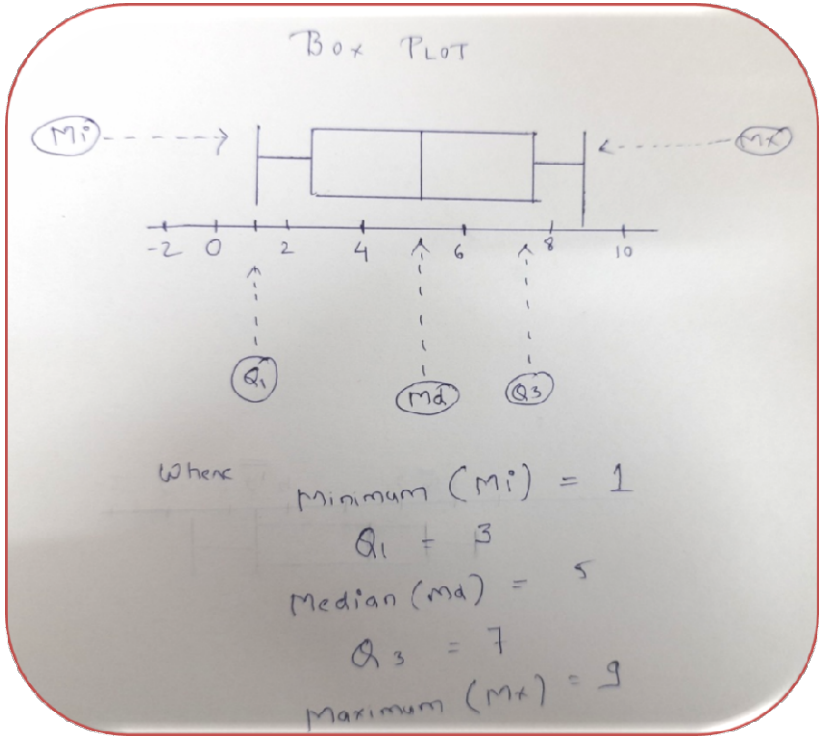
FIVE NUMBER SUMMARY: CONCLUSION

Solution: From above solutions, now we can figure out following points

Data set: after removal of Outliers

1	2	2	2	3	3	4	5	5	5	6	6	6	6	7	8	8	9

Minimum	→	1
First Quartile (Q1)	→	3
Median	→	5
Third Quartile (Q3)	→	7
Maximum	→	9



Note: In above diagram, Mi, Mx and Md is used for illustrative purpose only .

INTERVIEW QUESTION: WHAT IS THE USE OF BOX PLOT?

Box Plot: It is used for the determination of Outlier.
: It gives u the visualization way, where the outlier is actually presents

INTERVIEW QUESTION: How BOX PLOT uses for the removal of Outlier?

With the help of Five Number Summary, like Minimum, Q1, Median, Q3 and maximum, we can identify Upper fence and lower Fence of Data Set and then with the help of Box Plot, we can remove Outliers