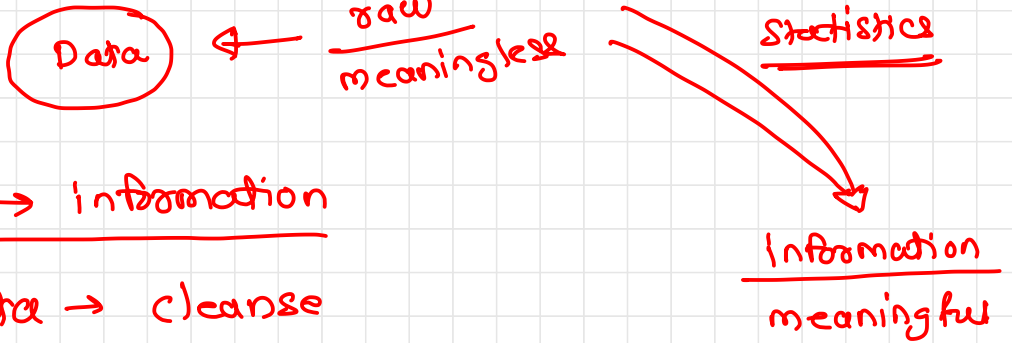
Statistics



raw → information

dirty data → cleanse

statistics

- collecting data : survey, advertisement
- organizing data : files (text, json, xml..), database → SQL / NoSQL
- displaying data : visualization of data → Graphs (scatter, line, piechart)
- analysis of data : understanding the data
- interpretation of data : ——
- future prediction : regression → classification

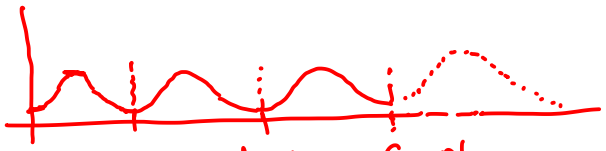
Introduction

- Discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data
- A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data
- Statistics has two meanings, as in plural sense and in singular sense
  - In plural sense, it means a systematic collection of numerical facts
  - In singular sense, it is the science of collecting, classifying and using statistics (data)
- Is both science of uncertainty and technology of extracting information from data
  - → tools (formula / methods
- Is used help us to make the decisions

Stages of investigation

[29, 29.2, 29.5, 30, 31 ... 26, 25.25]
temperature prediction : past temperature (data) 26 → 36

- Collection of data
  - It is the first stage of investigation and is regarding collection of data
  - It is determined that which method of collection is needed in this problem and then data are collected
- Organization of data
  - The data are simplified and made comparative and are classified according to time and place
    - store : csv, json, table, database
- Presentation of data
  - Organised data is made simple and attractive : Graphs
  - The data presented in the form of tables diagrams and graphs

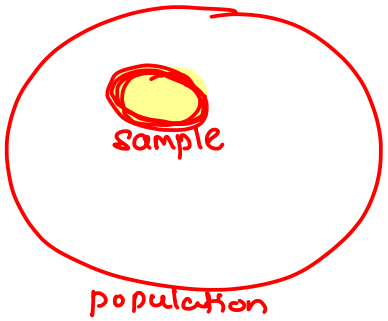


- Analysis of data
  - To get correct results, analysis is necessary (→ mandatory) ⇒ formula ⇒ method ⇒ algorithm
  - It is often undertaken using Measures of central tendencies, Measures of dispersion, correlation, regression and interpolation etc
    - Measures of central tendencies → mean / mode / median
    - Measures of dispersion → range, quartiles, IQR, variance, std devian, mean deviation
- Interpretation of data
  - In this last stage, conclusions are enacted
  - Use of comparisons is made
  - On this basis, forecasting is made

Population and Sample

1 Population

- A group that has been designated for gathering the data from
- Population is always defined first, before starting the data collection process for any statistical study
- Every member is considered
- It is impossible to contact every member if the population is very large
- Represented by **N**

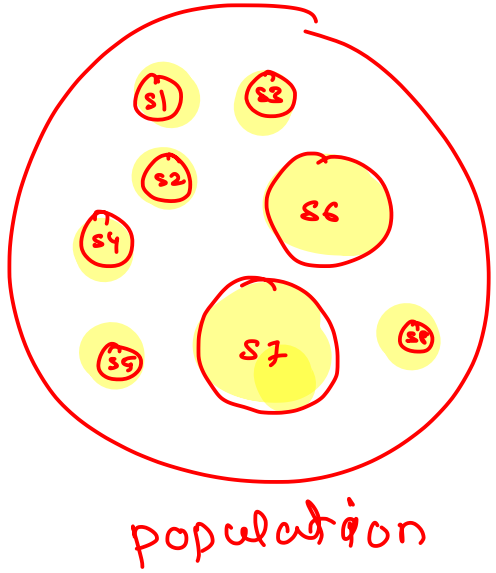


2 Sample

- It is the part of population which is selected randomly for the study
- Subset or small portion of population
- The sample should be selected such that it represents all the characteristics of the population
- The process of selecting the subset from the population is called **sampling** and the subset selected is called the **sample**
- Represented by **n**

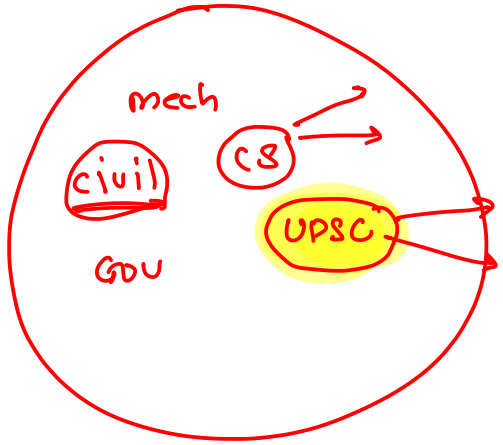
Sampling

- Process of getting a subset out of the population
- One of the great benefits of sampling is that a reasonably sized random sample will almost always reflect the population
- What is the challenge then ?
  - How to avoid bias?
  - How to choose sample?



Sampling Bias

- A bias in which samples are collected in such a way that some members of the intended population have a lower sampling probability than others
- It results in a **biased sample**, a non-random sample of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected
- If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling
- Types
  1. Selection Bias
  2. Survivorship Bias



Selection Bias

- Favors those members of a population who are more inclined and able to answer the questions
- Types
  - Under-coverage Bias
    - Making too few observations or omitting entire segment of a population
    - E.g. hospital survey of employee conducted during daytime hours. Neglects the employees working in night shift
  - Self-selection Bias
    - People who volunteer may differ significantly from those in the population who don't
    - E.g. an online survey about sports team. Only people who feel strongly about the team will answer the survey
  - Healthy-user Bias
    - The sample may come from a healthier segment of the overall population – people who walk/jog, work outside, follow healthier behaviours, etc.
    - E.g. Polling customers at a fruit stand to study a connection between diet and health. Those polled likely do other things that have greater impact on their health

Survivorship Bias

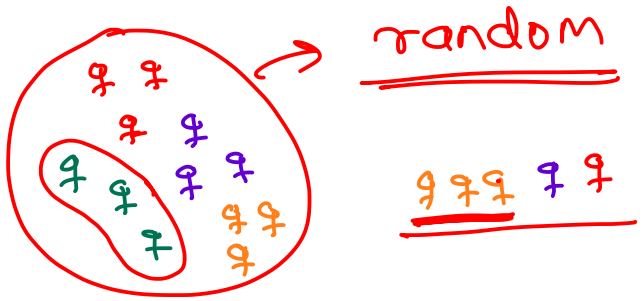
- If a population improves over time, it may be due to lesser members leaving the population due to death, expulsion, relocation, etc.

Sampling methods

- Describes how the sampling can be carried out
- Types
  - Radom sampling
  - Stratified Random sampling
  - Cluster sampling

Radom sampling

- Random sampling means every member of a population has an equal chance of being selected
- However, since samples are usually much smaller than populations, there's a chance that entire demographics might be missed



Stratified random sampling

- Stratified random sampling ensures that groups within a population are adequately represented
- First, divide the population into segments based on some characteristic [grouping]
- Members cannot belong to two groups at once [mutually exclusive]
- Next, take random samples from each group
- The size of each sample is based on the size of the group relative to the population
- E.g.
  - A company wants to conduct a survey of customer satisfaction
  - They can only survey 10% of their customers
  - They want to ensure that every age group is fairly represented
  - To obtain a 10% sample, take 10% from each group

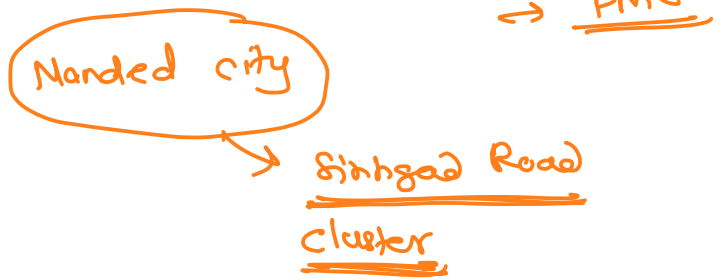
20-29	30-39	40-49	50-59	Total
1400	4450	3200	950	10000

→

20-29	30-39	40-49	50-59	Total
140	445	320	95	1000

Cluster sampling

- The idea is to break the population down into groups and sample a random selection of groups, or clusters.
- Usually this is done to reduce costs.
- E.g. Surveying only handful of neighborhoods

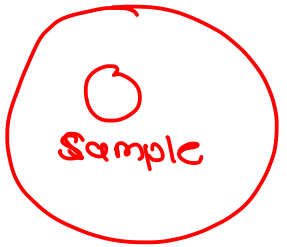


* is Road condition in Pune Good ?

→ population : ① people living in Pune, : transport authority, students, workers, ...
② visitors visiting Pune
③ drives
④ Punekars : students, workers

Parameter Vs Statistics

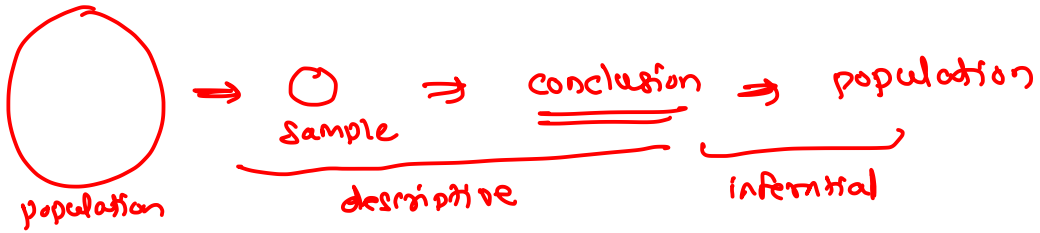
Parameter	Statistics
Is a measure that describes entire population	Is a measure that describes only a sample of population
E.g. - Actual voter turnout - Age of every person in country	E.g. - The portion of people in opinion polls - Age of people in sample



Statistics types

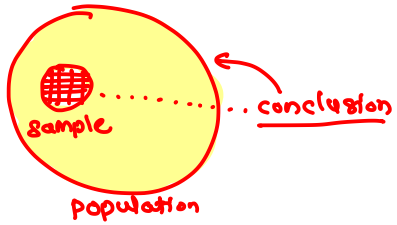
- **Descriptive Statistics**
  - It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc
  - Data can be summarized and represented in an accurate way using charts, tables and graphs

- **Inferential Statistics**
  - It is about using data from sample and then making inferences about the larger population from which the sample is drawn
  - The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population
  - It determines the probability of the characteristics of the sample using probability theory
  - The most common methodologies used are hypothesis tests, Analysis of variance etc



Statistics types

Descriptive	Inferential
Concerned with describing the target population	Make inferences from sample and generalize them for the entire population
Organize, analyze and present the data in meaningful manner	Compare, test and predict the future outcomes
Final results are shown in the form of charts, tables or graphs	Final result is probability score
Describes the data already known	Tries to make the conclusion about the population that is beyond the data available
Tools: - Measures of central tendency - Measures of dispersion	Tools - Hypothesis testing - Analysis of variance



Individuals and Variables

- Meaning outside the statistics
- Individuals
  - Are the people
  - E.g. 60 students for the course
- Variable
  - Is the factor that can vary
  - E.g. time the shop can make cake

- Meaning inside the statistics
- Individuals → sample data
  - Objects included in the study
  - E.g. records, people, objects
- Variable
  - Characteristics of individuals to be measured or observed
  - E.g. age or name of the person

Individuals and Variables

variable : Random variable / feature

Sample Data

	Name	Email	Phone	Address	Age
	Qualitative	Qualita...	Quant...	Qualita...	Quant...
1	Person1	p1@test.com	7845343456	Pune	56
2	Person2	p2@test.com	6644664466	Mumbai	67
3	Person3	p3@test.com	7676767676	Satara	30
4	Person4	p4@test.com	8989898989	Karad	50

← individual / record / object / item / row

↑ variable / characteristic / attribute / column

Type of Data

Type of data

continous

categorical

factors in R

(levels)

eg: city: Pune, mum, pune, satara ...

colors: red, brown, red, blue, green

Quantitative (Numeric)

20-29

interval

50-60 → 1

60-70 → 2

70-80 → 3

interval

continuous

→ 120kg

{119-121}

120

ratio

discrete

→ 90%.

Qualitative (textual)

Nominal

- Not ordered
- red, blue, green

Ordinal

→ values are ordered

→ unsatisfactory, satisfactory, good, excellent

→ grade → A, B, C, D

1

Variables: Qualitative

- Variables that are not measurement variables
- Also known as categorical variables
- Take category or label values and place an individual into one of several groups
- Each observation can be placed in only one category
- The categories are mutually exclusive
- E.g. political party, profession, gender, whether person smokes
- Types
  - Nominal
    - Can not be ordered from smallest to largest
  - Ordinal
    - Can be arranged in order of categories
    - But difference between data values can no be calculated or meaningless

Variables: Quantitative

- Variables whose values result from counting or measuring something
- Also known as continuous variable
- Take numerical values and represent some kind of measurement
- E.g. height, weight, age
- Types
  - Interval
    - There is no true zero
  - Ratio
    - There is a true zero

Example

- Identify individuals, variables, and categorical variables from the data set

Biscuits	Type	Calories	Carbs
Krack Jack	Sweet and Salty	16	2
Monaco	Salty	15	1.8
Marie	Sweet	25	6
Oreo	Sweet	53	8
Nice	Sweet	39	5

Series

- Collection of data points is called as a series
- Types
  - Time Series
    - A series of data that is arranged chronologically, or in relation to time
  - Frequency Series
    - A series of data that is formed along with the frequencies of their occurrences
    - Types
      - Individual series
      - Discrete series
      - Continuous series

Individual Series

- Each value of the variable occurs for only once
- The frequency of occurrence of all the values in such a series is only one
- Such series are displayed without the frequency column
- E.g.
  - Marks: 40 60 80 80 45 85 90 67
  - Ages: 30 40 35 45 56 60

Age	Frequency
30	1
40	1
35	1
45	1
56	1
60	1

Discrete series

- The different values of a variable are shown in a discontinuous manner along with their respective frequencies
- Such a series can also be arranged either in ascending, or in descending order
- E.g.

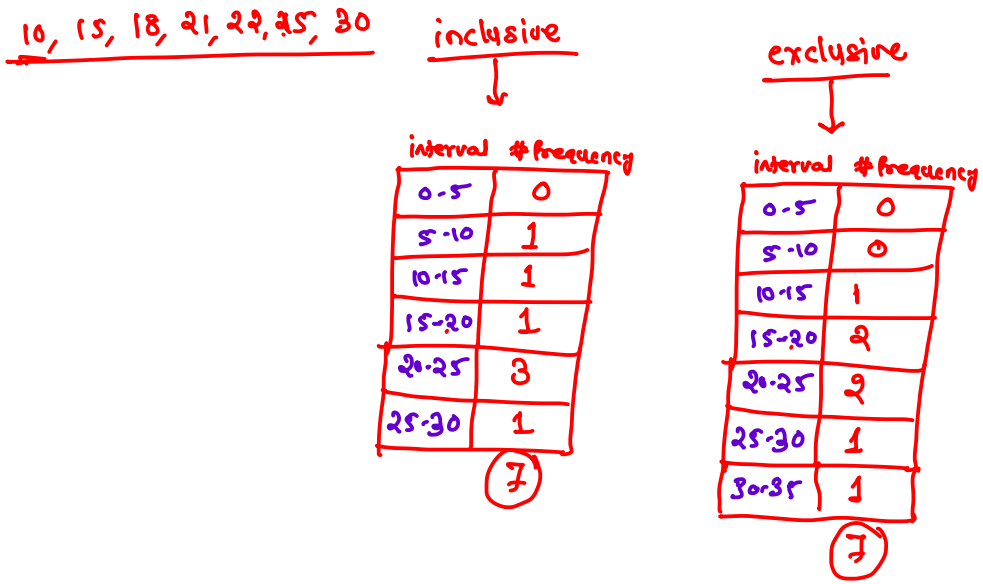
Marks	# Students
80	2
40	4
45	3
60	6
86	3
90	3
70	10

Wages	# Workers
150	10
160	20
170	10
200	6
500	4
1000	2
1500	1

80, 80, 40, 40, 40, 40 . . . .

Continuous Series

- Different values of the variables are stated in a continuous manner along with their respective frequencies
- Such series can be stated either in the form exclusive, or in the from of inclusive class intervals along with their respective class frequencies
- E.g.



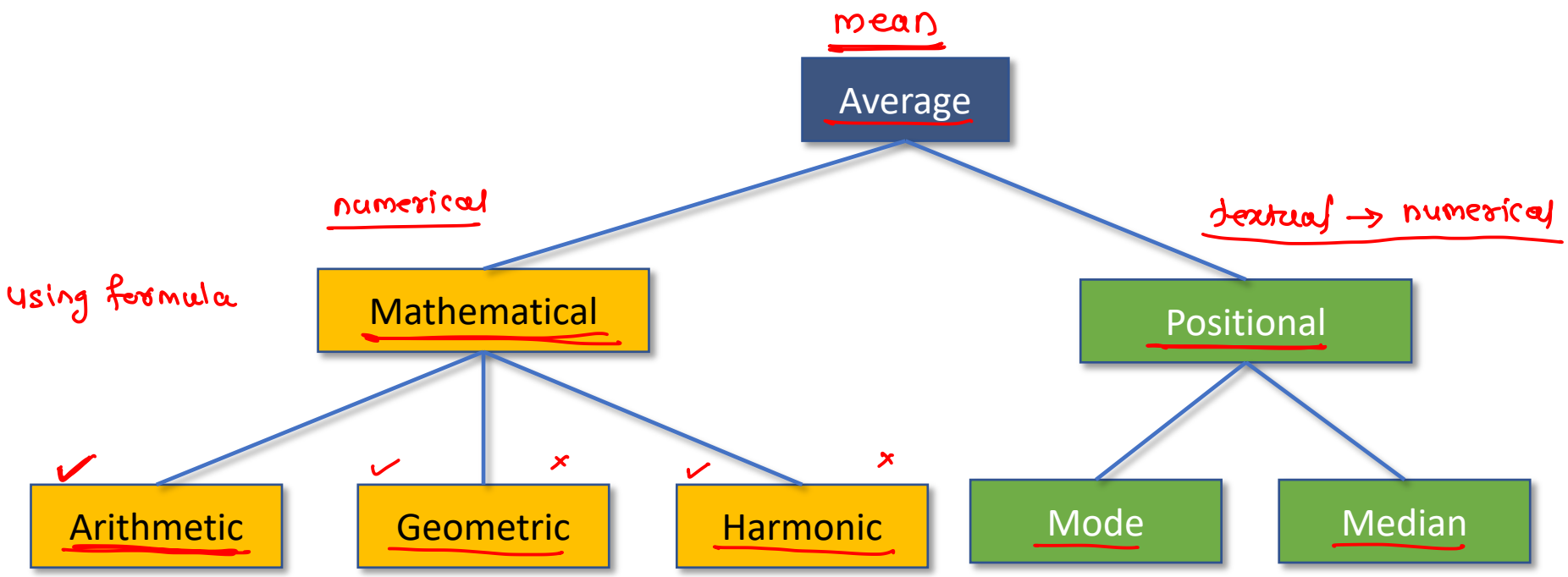
Marks Range	# Students
10-20	5
20-30	0
30-40	20
40-50	15
50-60	10
60-70	15
70-80	10
80-90	20
90-100	5

Measures of Central Tendency

Introduction

- Is a summary statistic that represents the center point or typical value of a dataset
- These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution

40, 50, 55, 60, 70, 45, 30, 48, 42



Example 1

Find out the mean, mode and median of the following data set

23, 30, 21, 32, 24, 22, 33, 21, 25

$$\text{mean} = \frac{\Sigma x}{n} = \frac{23+30+21+32+24+22+33+21+25}{9}$$

$$= \frac{231}{9} = 25.6$$

mean = 25.6

mode : most repeated value

[12, 15, 26, 12, 26, 12]

12	3
15	1
26	2

mode = 12

single mode

[12, 15, 26, 12, 26, 15, 30]

12	2
15	2
26	2
30	1

mode = [12, 15, 26]

Multi-mode

individual series

[12, 15, 26, 17, 18, 20]

12	1
15	1
26	1
17	1
18	1
20	1

mode

- No mode ✓
- [12, 15, 26, 17, 18, 20] ↑

Median : middlemost value

odd no of values

[12, 15, 30, 21, 22, 30, 50]

① arrange the values in asc order

[12, 15, 21, 22, 30, 30, 50]

↑ median

median = 22

even no of values

[12, 15, 30, 22, 21, 40]

① arrange values in asc order

[12, 15, 21, 22, 30, 40]

average

$$\text{median} = \frac{21+22}{2} = \frac{43}{2} = 21.5$$

median = 21.5

① [10, 20, 15, 20, 21, 30, 35]

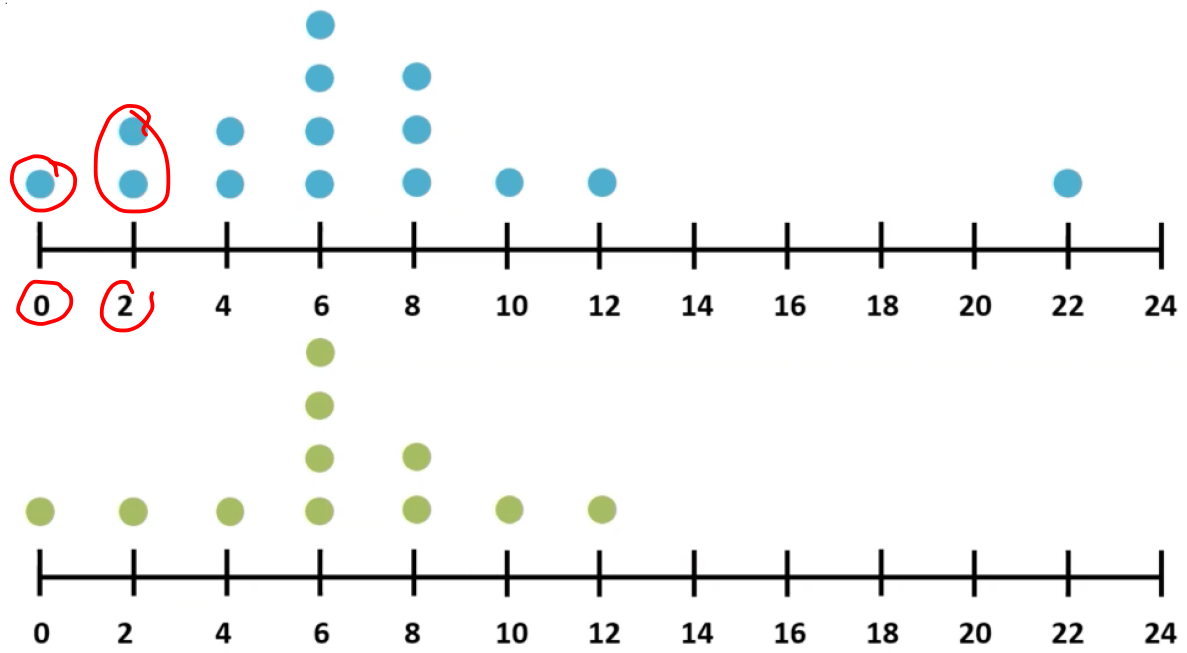
② [30, 12, 15, 18, 21, 5, 15, 9]

③ [15, 18, 30, 15, 18, 30, 5]

Example 2

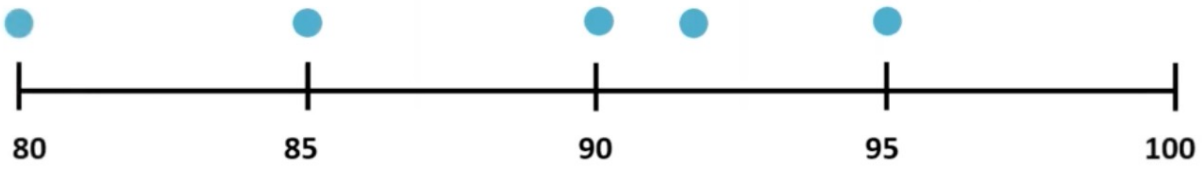
James interviewed Juniors and Seniors at his university, asking them how many pieces of fruit they eat each day. The results are shown below:

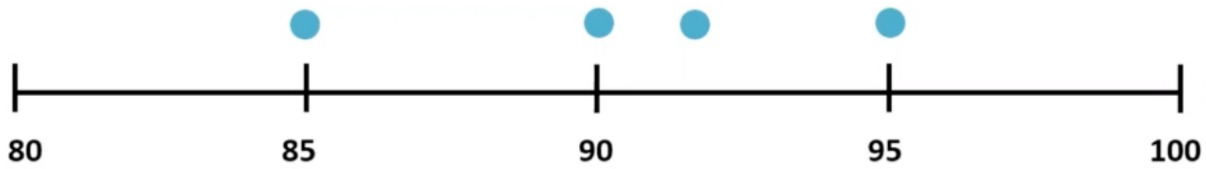
- Which group has high mean?
- Is mean a good measure for centre of the distribution for seniors? Can we use mean for centre of the distribution for Juniors?



Example 3

- Atharv played 5 rounds of golf, and his lowest score was an 80. The scores of the first 4 rounds and the lowest round are show in the following dot plot
- It was found that Atharv broke some rules when he scored 80, so that score will be removed from the data set





- How will the removal of the lowest score affect the mean and the median?
  - Both the mean and the median will decrease , but the mean will decrease by more than the median.
  - Both the mean and the median will decrease, but the median will decrease by more than the mean.
  - Both the mean and the median will increase, but the mean will increase by more than the median.
  - Both the mean and the median will increase, but the mean will increase by more than the median.