Statistical Foundations Notes and Examples

Statistical Learning

Statistics include numerical facts and figures. For instance:

- The largest earthquake measured 9.2 on the Richter scale.
- Men are at least 10 times more likely than women to commit murder.

The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on **how the numbers are chosen and how the statistics are interpreted**. You will find that the numbers may be right, but the interpretation may be wronged

E.g. A **new advertisement** for Ben and Jerry's ice cream introduced in late May of last year resulted in **a 30% increase in ice cream sales** for the following three months. Thus, **the advertisement was effective**!! Can we conclude that?

A **major flaw** is that ice cream consumption **generally increases in the months of June, July, and August** regardless of advertisements. This effect is called a **history effect** and leads people to interpret outcomes as the result of one variable when another variable (in this case, one having to do with the passage of time) is actually responsible

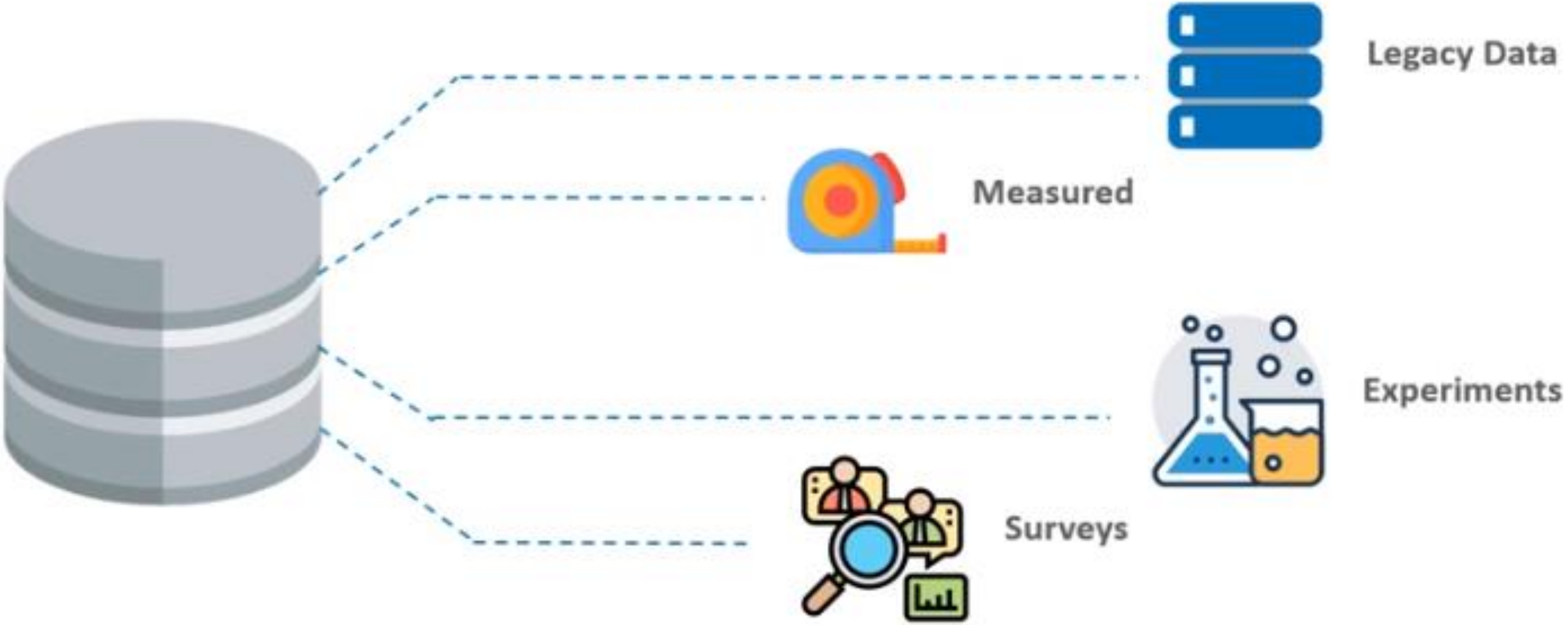
- As a whole, these examples show that statistics are ***not only facts and figures***; they are something more than that. In the broadest sense, “statistics” refers to **a range of techniques and procedures** for analyzing, interpreting, displaying, and making decisions based on data.**!**

Statistical Learning v/s Machine Learning

- Both methods are data dependent. However, Statistical Learning relies on rule-based programming; it is formalized in the form of **relationship between variables**, where Machine Learning **learns from data** without explicitly programmed instructions.
- Statistical Learning is based on a **smaller dataset with a few attributes**, compared to Machine Learning where it can **learn from billions of observations and attributes.**
- Statistical Learning is **mostly about inferences**, most of the idea is generated from the sample, population, and hypothesis, in comparison to Machine Learning which **emphasizes predictions, supervised learning, unsupervised learning, and semi-supervised learning**.
- Statistical Learning is **math intensive** which is based on the coefficient estimator and requires a good understanding of your data. On the other hand, Machine Learning **identifies patterns from your dataset** through the iterations which require a way less of human effort.

Data in Statistical Analysis

Data is a set of recorded facts, numbers, or events *with no meaning*



Basic Classification of Data

- **Categorical data** represents types of data that can be divided into groups – called Categories/Classes
- **Numerical data** represents data that can take values on a numeric scale

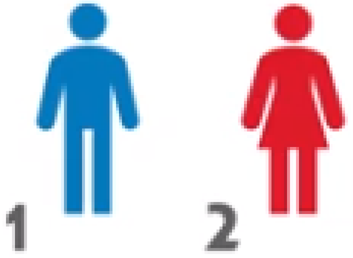
Categories of Colors

Vs.

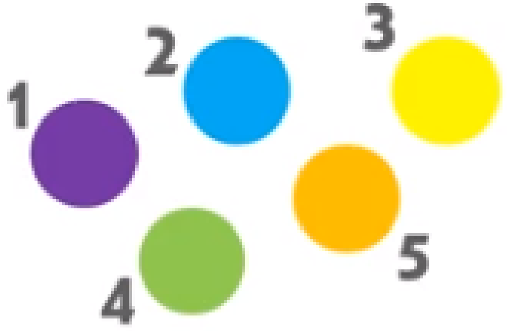
Gradient of Colors

Categorical Data – Nominal Data

Gender



Color



Food Preference

Veg 1

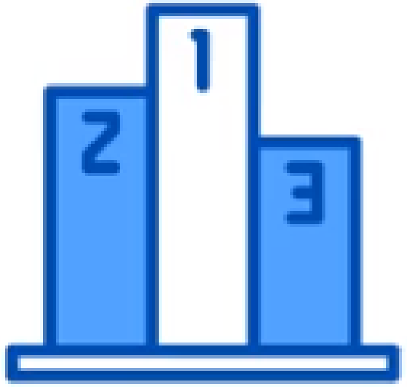
Non-veg 2

Sometimes, numbers are assigned to the categories, but they hold ***no mathematical significance or order***

Categorical Data – Ordinal Data

Ordinal data consists of data with an ordered series

Ranks



Restaurant Ratings

The levels may be in the form of numbers or labels, but they all denote an order

Numerical Data

Includes things that can be measured rather than classified or ordered

Days to deliver an order

Ages of people

Daily Temperature

Numerical Data – Discrete Data

Data for which only a finite number of distinct values are possible

Customer's Name	Shoe Size
Shahid Collister	9
Eric Hoffmann	7
Philip Fox	5
Frank Olsen	9
Mary O'Rourke	8

Shoe sizes

Intermediate values such as 5.1 or 7.3 do not have a meaning

Numerical Data – Continuous Data

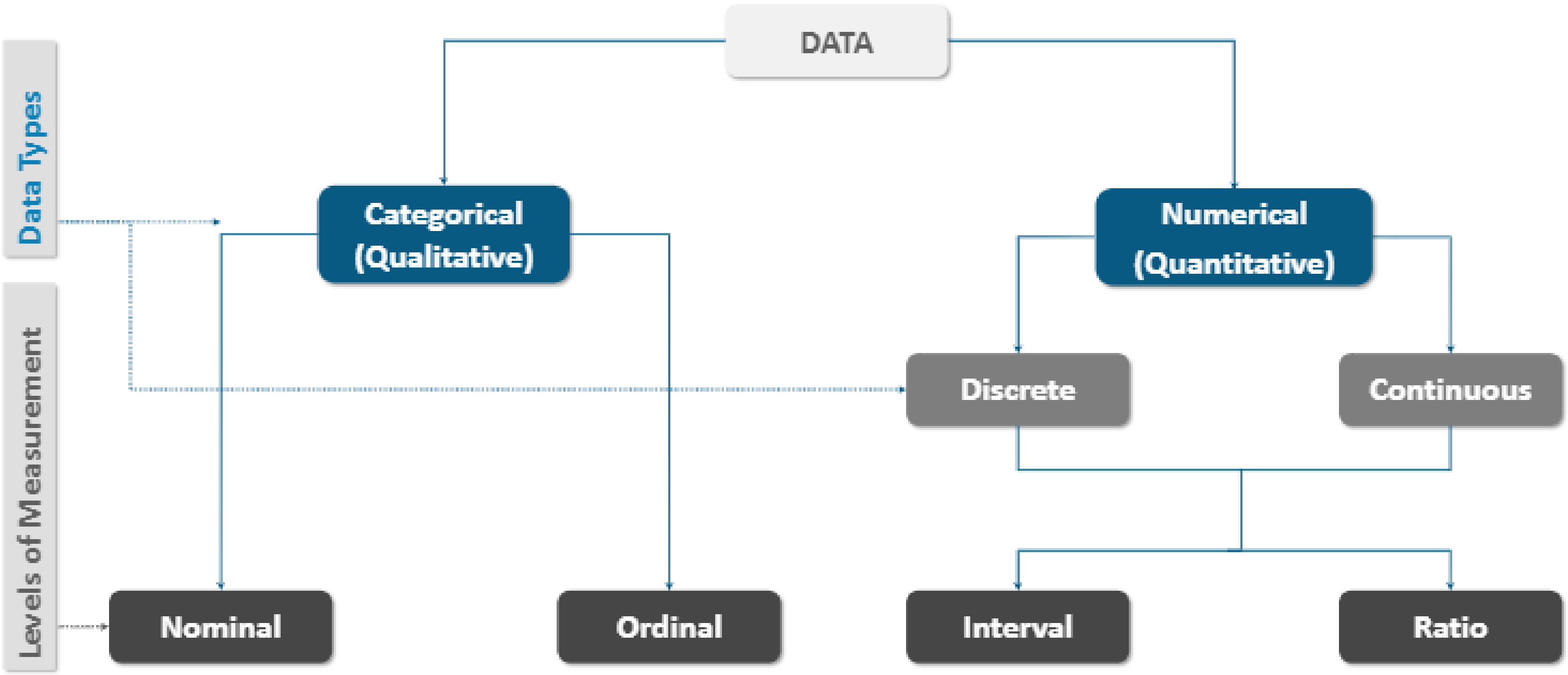
Data which can have almost any numeric value in real numbers

Patient ID	Weight in Kg.
12	86.5
15	91.3
11	56.1
7	70.9
5	60.34

Weights

Properties such as Mean, Median, and Standard Deviation are used to learn more about Numerical Data

Data Types: Summary



- Qualitative
  - Nominal (labels) – cities, colours, gender, marital status, countries – no order
  - Ordinal – grades, ratings, income groups
- Quantitative
  - Discrete (specific set of values) – dice , shoe size
  - Continuous (continuous) – weight, height

Levels of Measurement

- A **nominal scale**, provides a name or category for each object
- In an **ordinal scale**, the objects are ordered
- In an **interval scale**, the same difference at two places on the scale has the same meaning.
- A **ratio** scale, in addition to the above, has the same ratio at two places on the scale also carries the same meaning AND has a true zero.

Levels of Measurement

- Interval scale (no true zero) v/s Ratio scale
- 0 kg vs 0 degree Centigrade (0 kg actually means no weight whereas 0 degree does not mean no temperature, its just that the scale calls is 0 degrees)
- kg has true zero, so ratio scale
- Centigrade has no true zero, so interval scale

Interval Scales

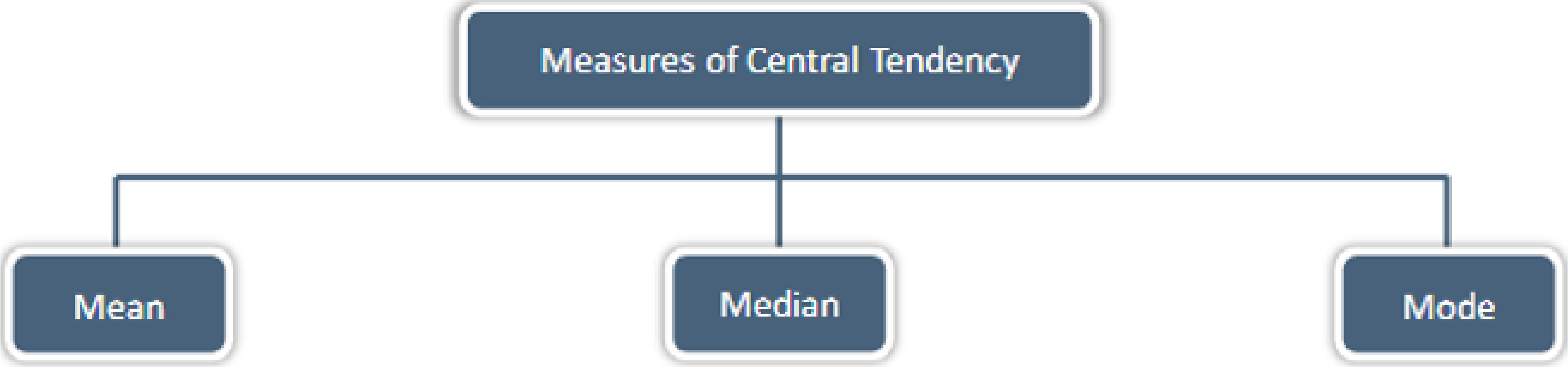
- Interval scales are numerical scales in which intervals have the same interpretation throughout.
- As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).
- Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name “zero.”

Ratio Scale

- The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured.
- You can think of a ratio scale as the three earlier scales rolled up in one.
    - Like a nominal scale, it provides a name or category for each object (the numbers serve as labels).
    - Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers).
    - Like an interval scale, the same difference at two places on the scale has the same meaning.
    - And in addition, the same ratio at two places on the scale also carries the same meaning.

Measures Of Central Tendency

- The data values for most numerical variables tend to group around a specific value
- **Measures of Central Tendency** describe to what extent this pattern holds for a specific variable
- Three commonly-used measures are :



Measures of Central Tendency

Mean

Median

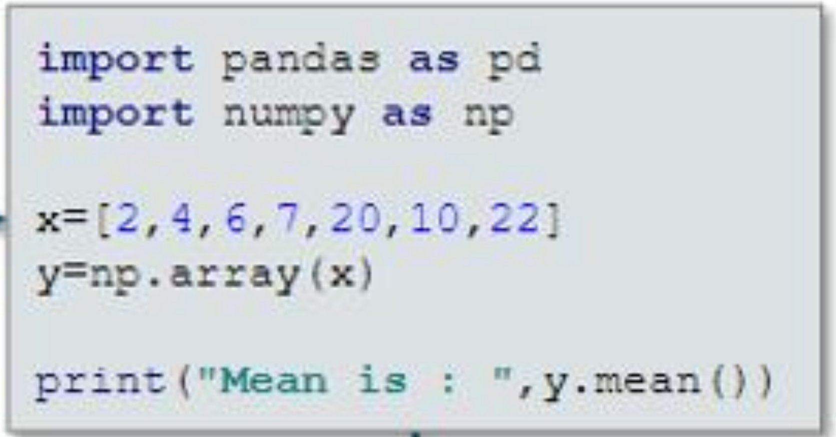
Mode

Mean

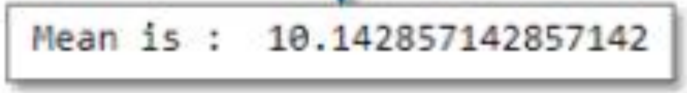
Represents the sum of all values in a dataset divided by the total number of the values

- The mean of a set of numbers x_1, x_2, x_3.....x_n is defined as,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



```
import pandas as pd
import numpy as np

x=[2,4,6,7,20,10,22]
y=np.array(x)

print("Mean is : ",y.mean())
```



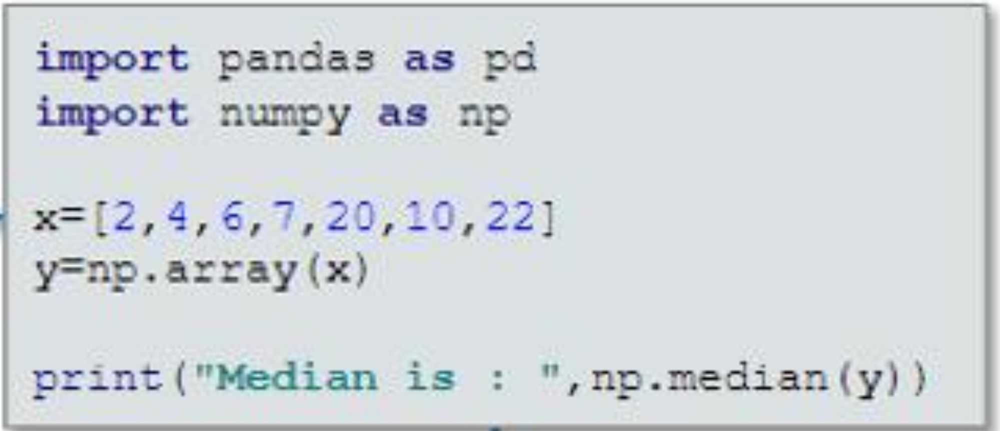
```
Mean is :  10.142857142857142
```

Median

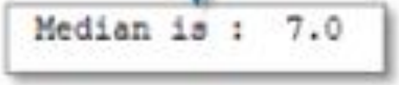
Represents the middle value in a dataset that is arranged in ascending order

- The statistical median of the data x_i is defined as,

$$\tilde{x} = \begin{cases} x'_{(n+1)/2}, & \text{if } n \text{ is odd;} \\ \frac{1}{2}\left(x'_{n/2} + x'_{1+n/2}\right), & \text{if } n \text{ is even.} \end{cases}$$



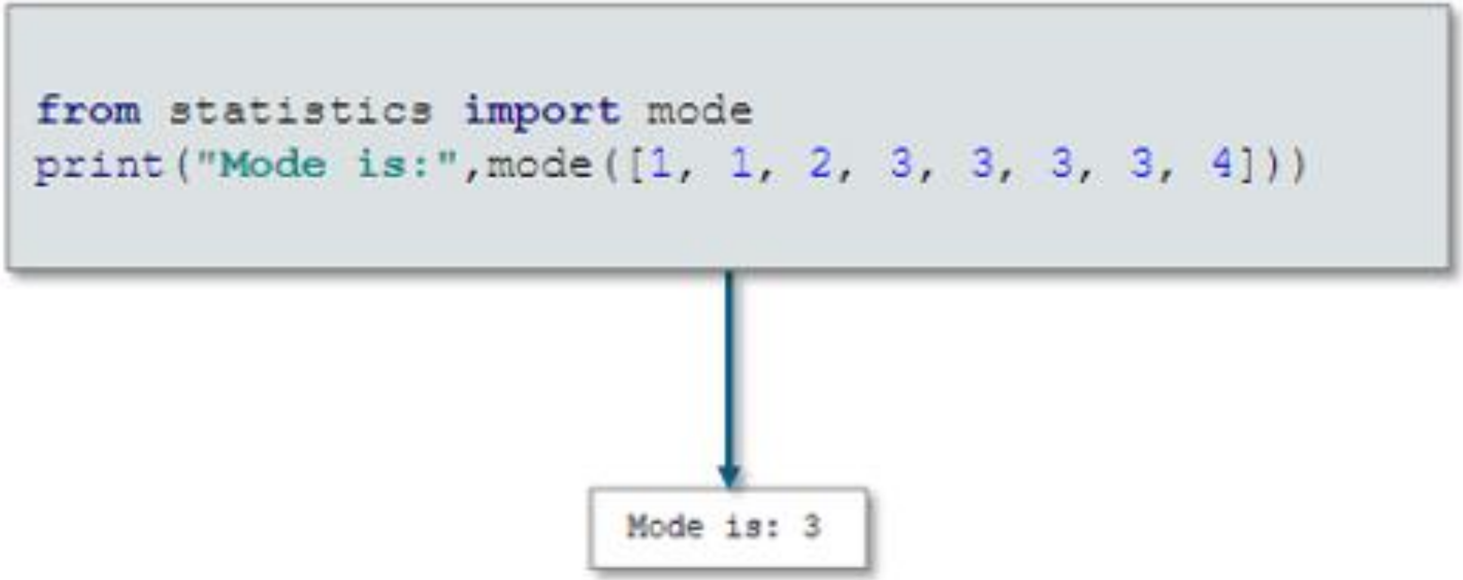
```
import pandas as pd
import numpy as np

x=[2,4,6,7,20,10,22]
y=np.array(x)

print("Median is : ",np.median(y))
```



```
Median is :  7.0
```

Mode

Defines the most frequently occurring value in a dataset



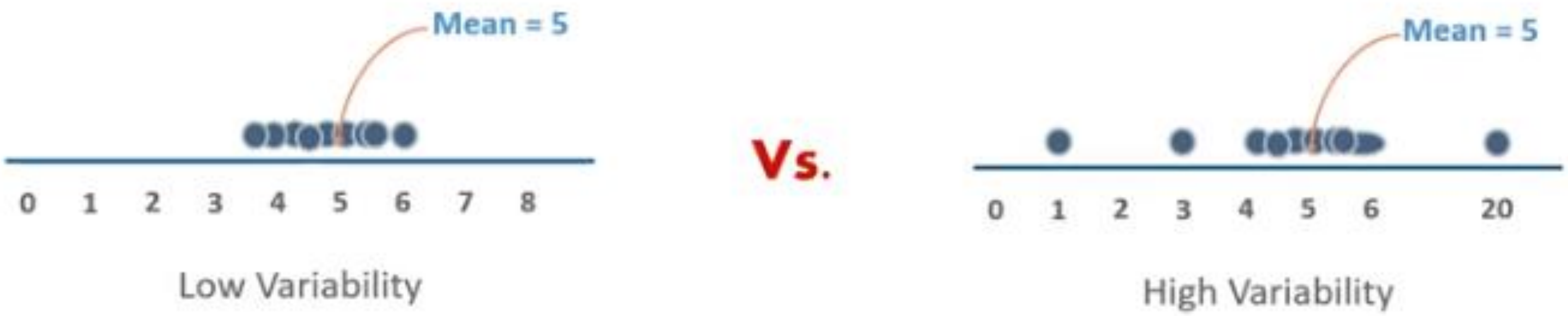
```
from statistics import mode
print("Mode is:",mode([1, 1, 2, 3, 3, 3, 3, 4]))
```

```
Mode is: 3
```

In some cases, a dataset may contain multiple modes while some datasets may not have any mode at all

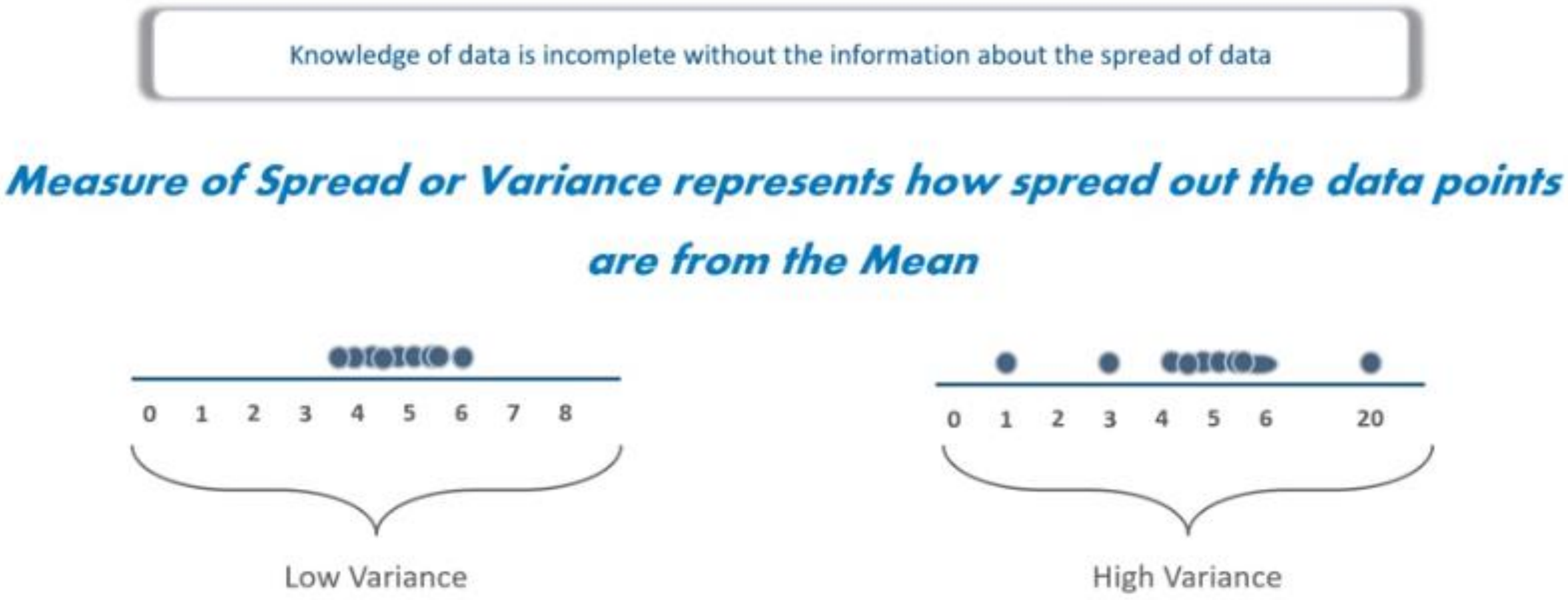
Measures of Spread

Measures of central tendency only represent the central location around which the data is present



Mean is same in this case irrespective of whether the data is closely packed or spread out

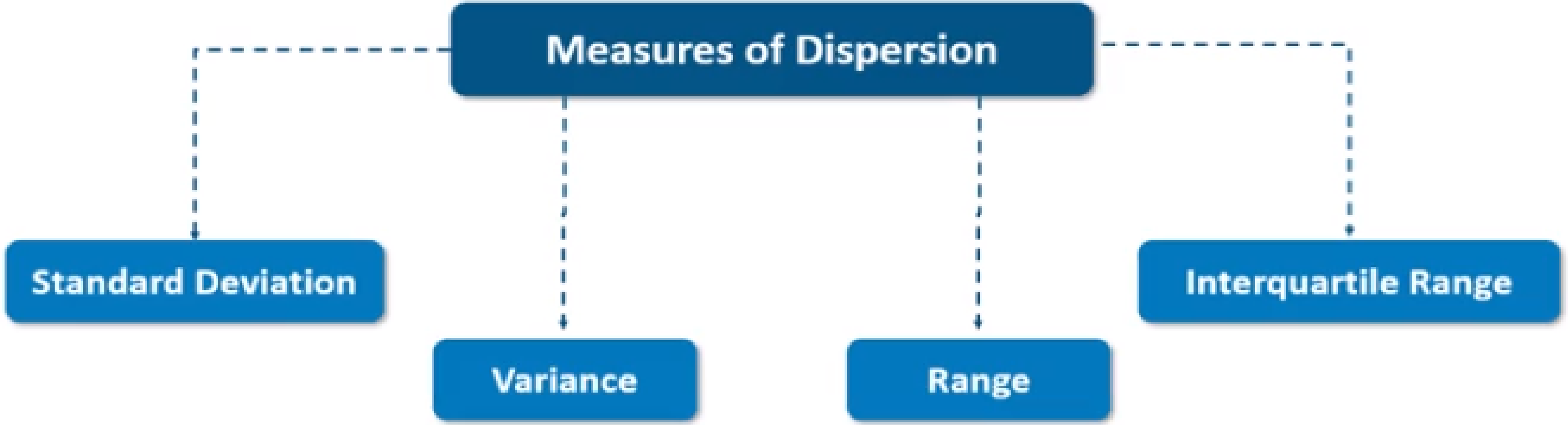
Measures of spread – why?



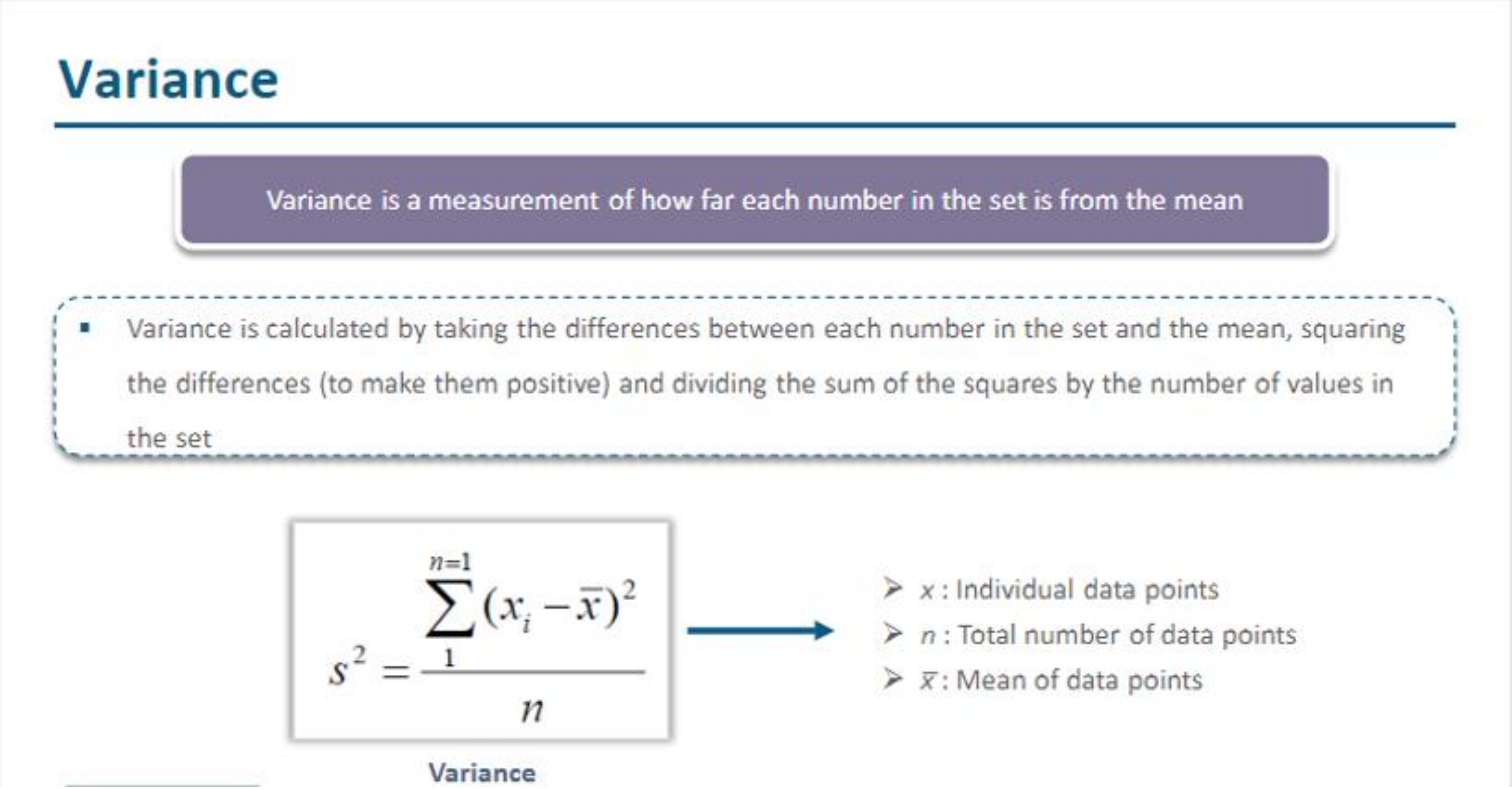
Knowledge of data is incomplete without the information about the spread of data

Measure of Spread or Variance represents how spread out the data points are from the Mean

Measures of Spread / Dispersion



- A **Measure of Dispersion**, or measure of spread, is used to describe the variability in a sample or population and how well the data is distributed
- It is usually used in conjunction with a measure of central tendency - mean or median, to provide an overall description of a set of data



Variance

Variance is a measurement of how far each number in the set is from the mean

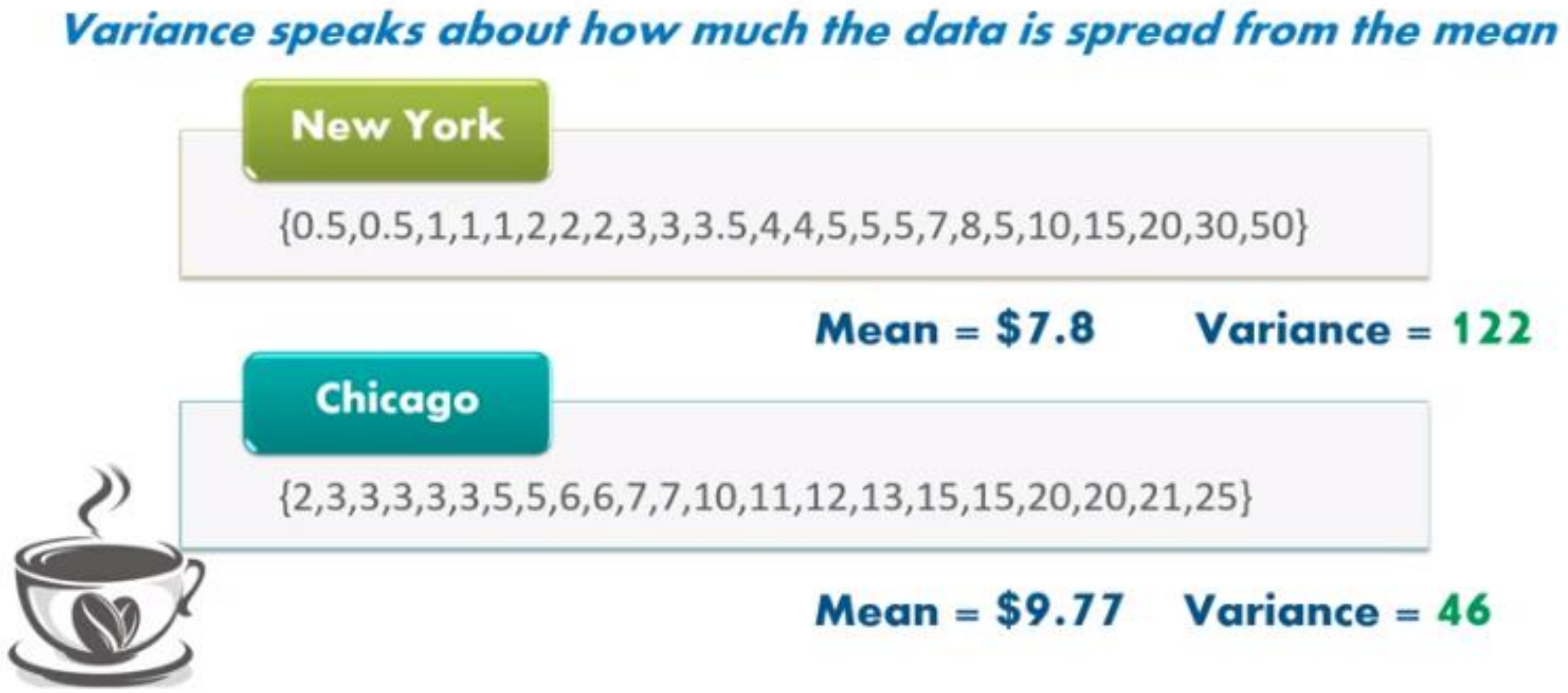
- Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set

$$s^2 = \frac{\sum_{1}^{n=1}(x_i - \bar{x})^2}{n}$$

Variance

- x : Individual data points
- n : Total number of data points
- $\bar{x}$: Mean of data points

Coffee price scenario - Variance



Variance speaks about how much the data is spread from the mean

New York

{0.5,0.5,1,1,1,2,2,2,3,3,3.5,4,4,5,5,5,7,8,5,10,15,20,30,50}

Mean = $7.8 **Variance = 122**

Chicago

{2,3,3,3,3,3,5,5,6,6,7,7,10,11,12,13,15,15,20,20,21,25}

Mean = $9.77 **Variance = 46**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

N : Size of the data
µ : Is the mean
x_i : Takes the data values

Standard Deviation

Standard Deviation is a measure of dispersion of a set of data from its mean

- If the data points are further from the mean, there is a higher deviation within the data set
- Standard deviation is calculated as the square root of variance by determining the variation between each data point relative to the mean, as shown below :

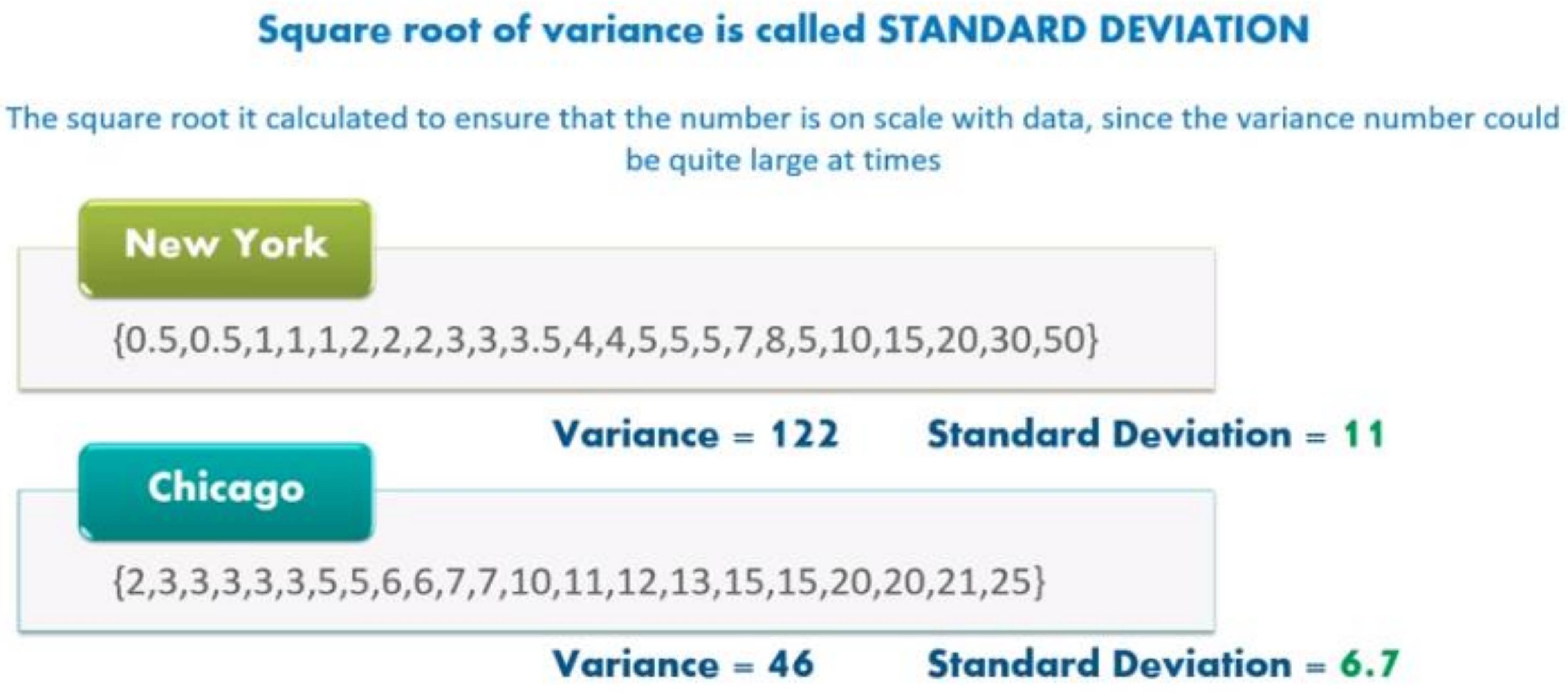
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{1}^{n=1} (x_i - \overline{x})^2}{n}}$$

→

- ➢ *x*: Individual data points
- ➢ *n*: Total number of data points
- ➢ $\overline{x}$: Mean of data points

Standard Deviation

Standard Deviation



Square root of variance is called STANDARD DEVIATION

The square root it calculated to ensure that the number is on scale with data, since the variance number could be quite large at times

New York

{0.5,0.5,1,1,1,2,2,2,3,3,3.5,4,4,5,5,5,7,8,5,10,15,20,30,50}

Variance = 122 **Standard Deviation = 11**

Chicago

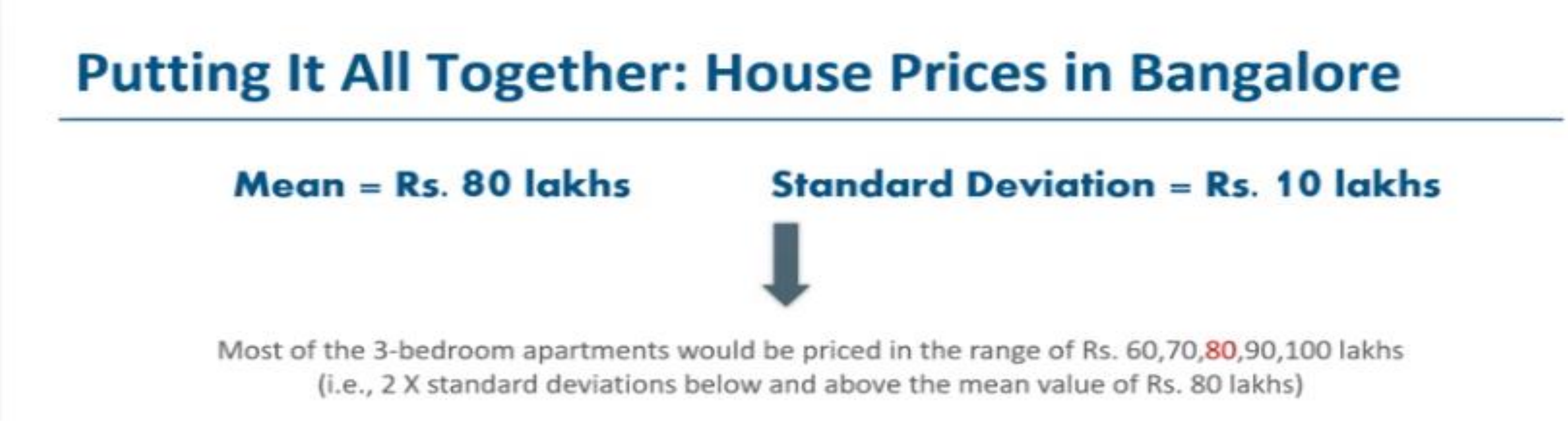
{2,3,3,3,3,3,5,5,6,6,7,7,10,11,12,13,15,15,20,20,21,25}

Variance = 46 **Standard Deviation = 6.7**

Putting It All Together: House Prices in Bangalore

Following is the data of the prices of 3-bedroom apartments in Bangalore (which is the silicon valley of India)

Mean	Max	Min
Rs. 80 lakhs (around 1 lakh USD)	Rs. 2 crores (around 2.7 lakh USD)	Rs. 50 lakhs (around 70000 USD)

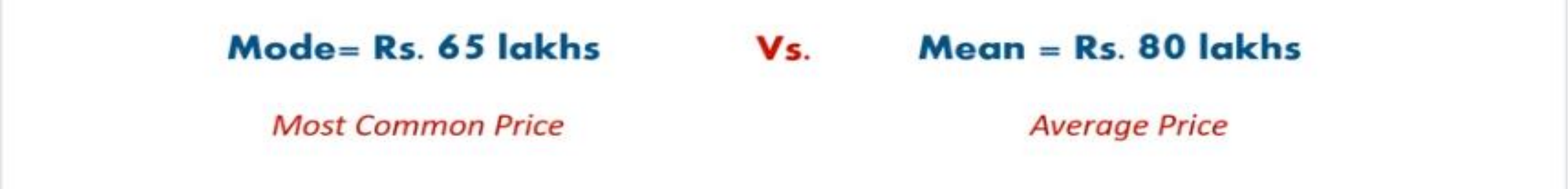


Putting It All Together: House Prices in Bangalore

Mean = Rs. 80 lakhs **Standard Deviation = Rs. 10 lakhs**

↓

Most of the 3-bedroom apartments would be priced in the range of Rs. 60,70,80,90,100 lakhs
(i.e., 2 X standard deviations below and above the mean value of Rs. 80 lakhs)



Mode= Rs. 65 lakhs **Vs.** **Mean = Rs. 80 lakhs**

Most Common Price *Average Price*

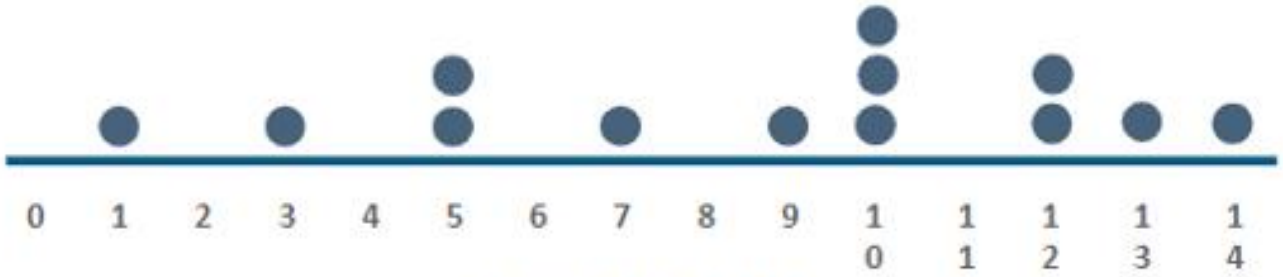
There could be a few high-priced houses which leads to such a scenario where MEAN is on the higher side

Range

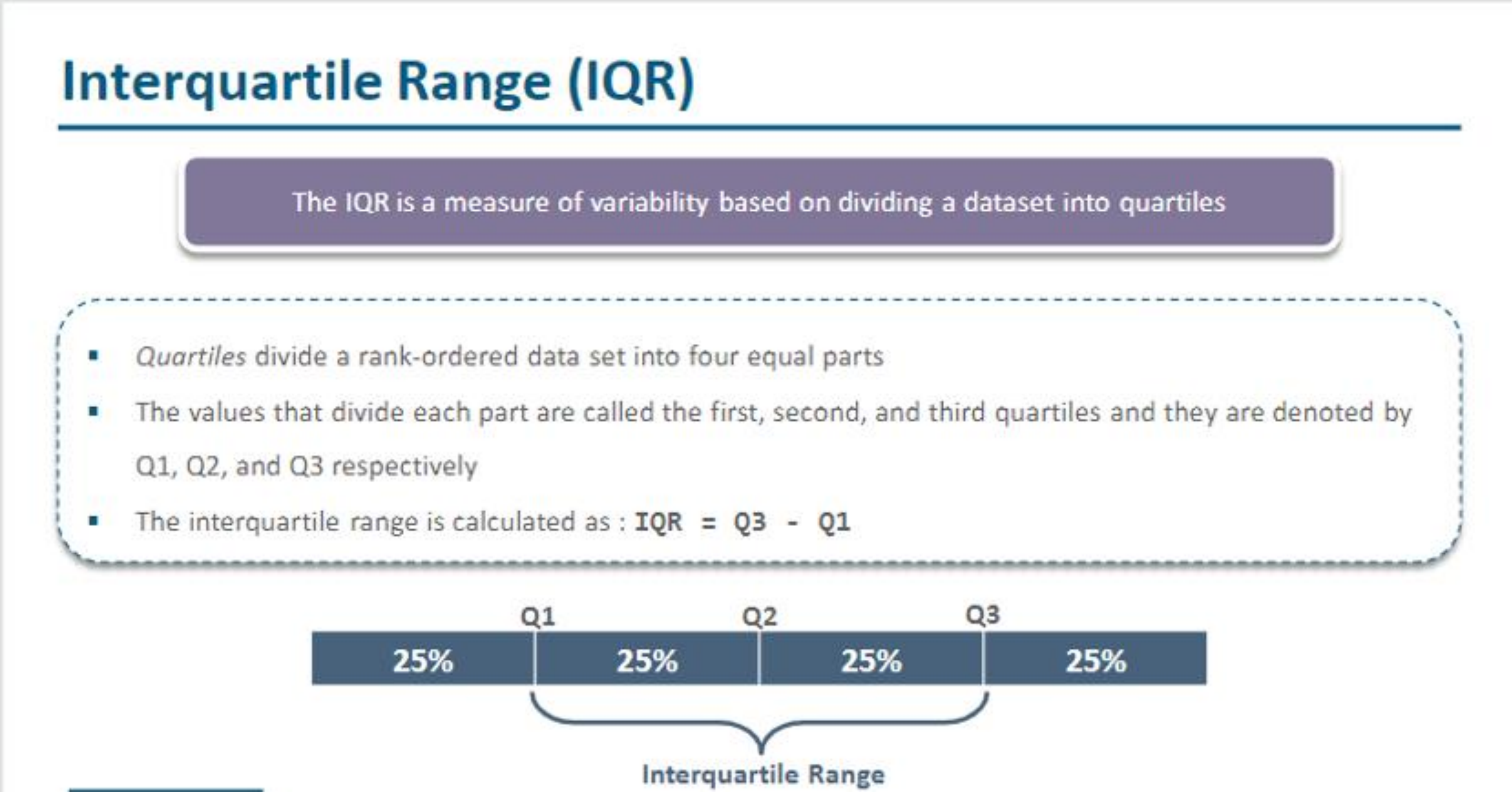
Range is the interval between the highest and the lowest value

- It gives a measure of how spread apart the values are
- It is simply calculated as :

  (maximum value – minimum value) = $\mathtt{Max}(x_i)\text{-}\mathtt{Min}(x_i)$



Range : 14 – 1 = 13



Interquartile Range (IQR)

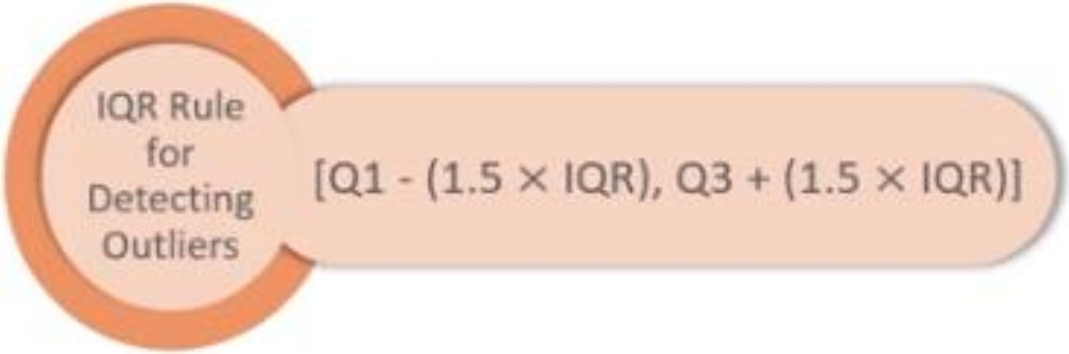
The IQR is a measure of variability based on dividing a dataset into quartiles

- *Quartiles* divide a rank-ordered data set into four equal parts
- The values that divide each part are called the first, second, and third quartiles and they are denoted by Q1, Q2, and Q3 respectively
- The interquartile range is calculated as : **`IQR = Q3 - Q1`**

Range, IQR, Outliers

Use an example of a ages of people in a group, to explain

- Range
- Q1 – lower quartile value
- Q3 – upper quartile value
- IQR – Inter quartile range
- Percentile
- Outliers :



IQR Rule for Detecting Outliers

[Q1 - (1.5 × IQR), Q3 + (1.5 × IQR)]

Population and Sample

Random Selection

Population

A collection or set of individuals or objects or events whose properties are to be analyzed

Sample

A subset of a population is called **Sample**. A well-chosen sample will contain most of the information about a population parameter

Why Use Sampling?

***Survey of India*, The National Survey and Mapping Organization of the country, wants to perform a survey about the eating habits of teenagers in India.**

Challenge: There are over **243** million teens in India!

Why Use Sampling?

EXPENSIVE

The estimated cost of AADHAR project (SSN equivalent of the USA) is close to Rs. 6000 Crores (close to a billion Dollars)

TIME CONSUMING

This project is going on since past 5-6 years and it is still an ongoing process

Note: The above points are only applicable when the population if very large and ROI on the project is not worthy

How Does an Incorrect Sample Look Like?

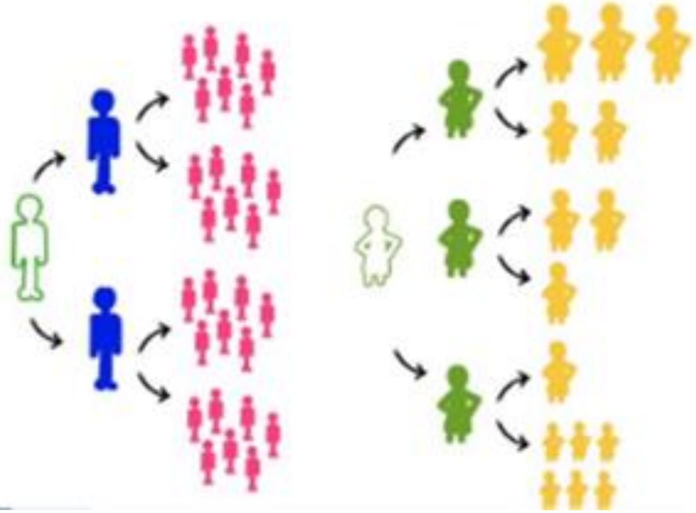
A collection of fruits

Random Selection

A sample from the collection of fruits

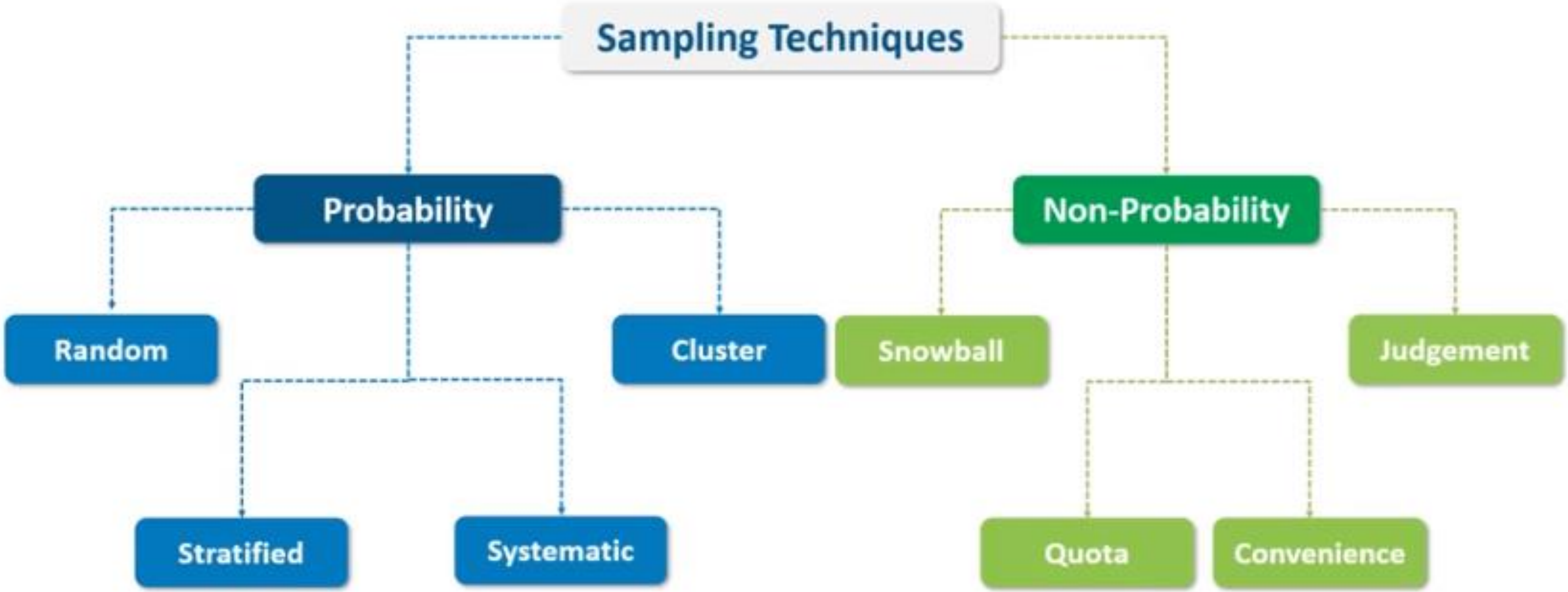
Sampling Techniques

In **Probability Sampling**, each member of the population has a known non-zero probability of being selected



In **Non-probability Sampling**, members are selected from the population in some non random manner

Different Types of Sampling Techniques



Random and Stratified are generally used – others not too useful

Random Sampling

Random Sampling

Stratified Sampling

Systematic Sampling

Cluster Sampling

The Probability of each member of the population to be chosen has an equal chance of being selected

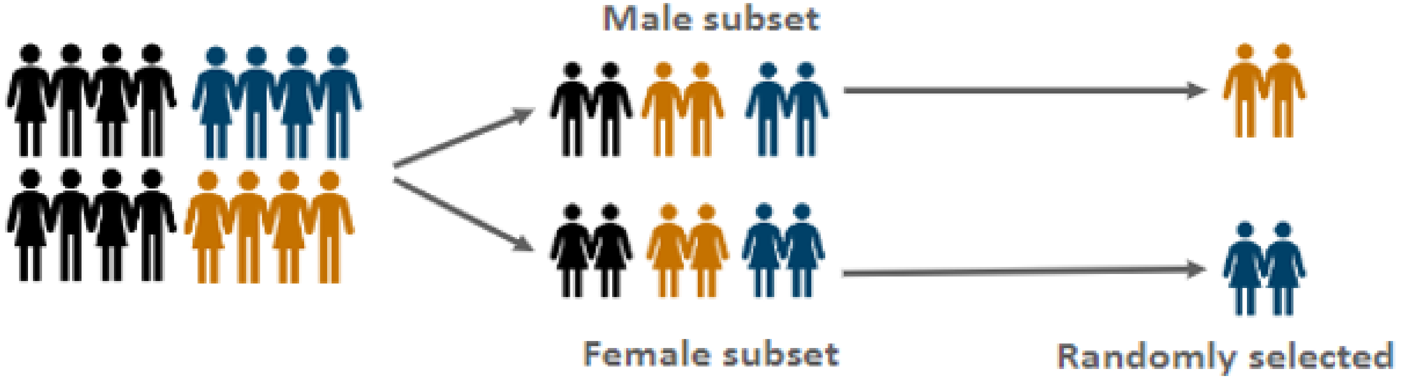
Stratified Sampling

Random Sampling

Stratified Sampling

Systematic Sampling

Cluster Sampling



- **Stratum:** Subset of the population sharing at least one common characteristic
- First step is to identify the relevant stratums and their actual representation
- Random sampling is then used to select subjects from each stratum

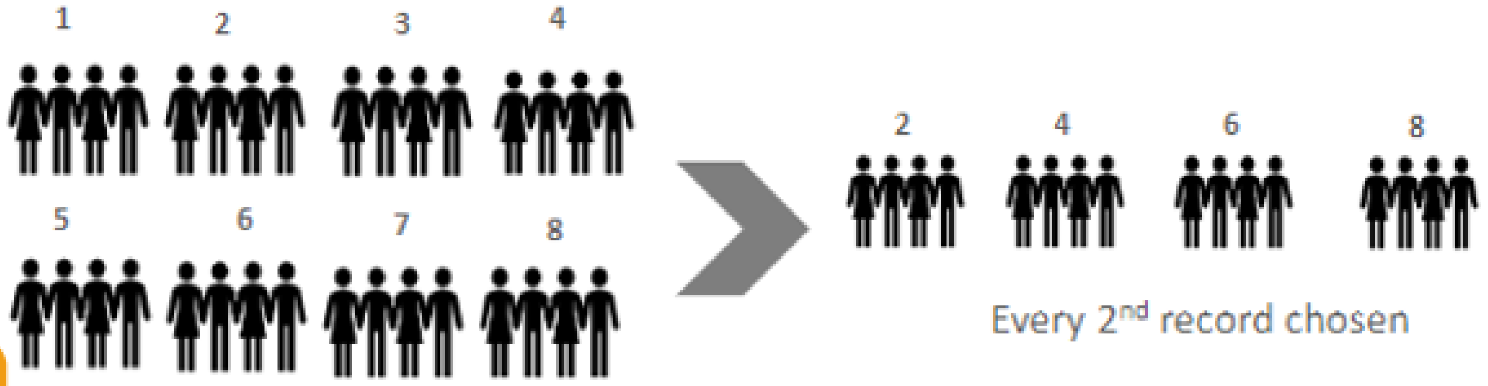
Systematic Sampling

Random Sampling

Stratified Sampling

Systematic Sampling

Cluster Sampling



Every nth record is chosen

Every 2nd record chosen

- Also known as **Nth name selection technique**
- Every Nth record is selected from a list of population
- Simpler than random selection technique

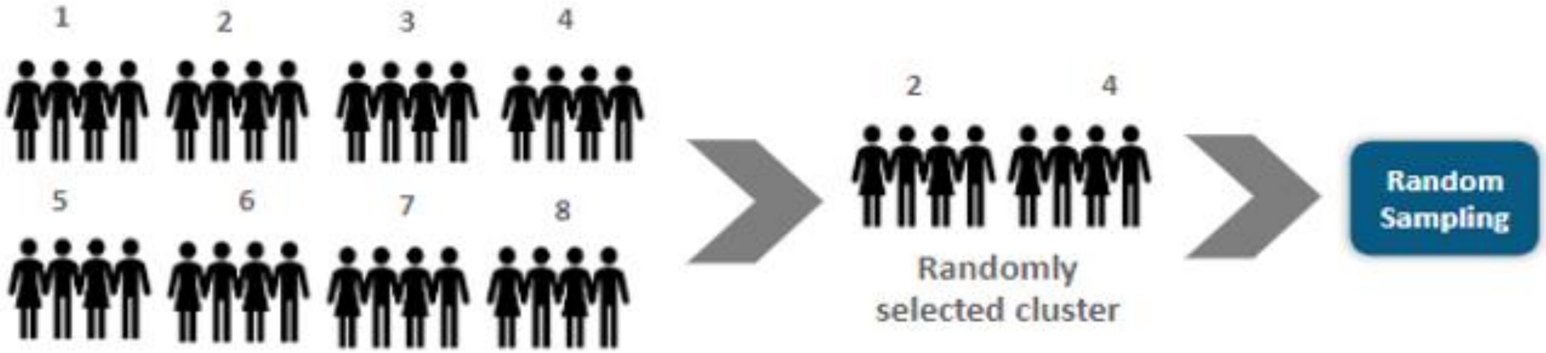
Cluster Sampling

Random Sampling

Stratified Sampling

Systematic Sampling

Cluster Sampling



- Randomly select a cluster from the population and perform simple random sampling on it
- **Example:** To select 1000 participants from the population of India, first randomly select a cluster such as a city or district and then select 1000 participants

Types of Non-probability Sampling

Snowball Sampling

It is a special nonprobability method used when the desired sample characteristic is rare

Quota Sampling

It is the Non-Probability equivalent of stratified sampling

Judgement Sampling

The researcher selects the sample based on judgment. This is usually an extension of convenience sampling

Convenience Sampling

Used in exploratory research for getting an inexpensive approximation of the truth