Introduction to Statistics

By

Prof(Dr) Premanand P Ghadekar

Outline

- What is Statistics
- Basic Terminologies in Statistics
- Sampling Techniques
- Types of Statistics
- Descriptive Statistics
- Inferential Statistics

What is Statistics

Statistics is an area of applied mathematics concerned with the data collection, analysis, interpretation and presentation.

Your Company has created a new drug that may cure Cancer. How would you conduct a test to confirm the drug effectiveness.
Covid vaccine

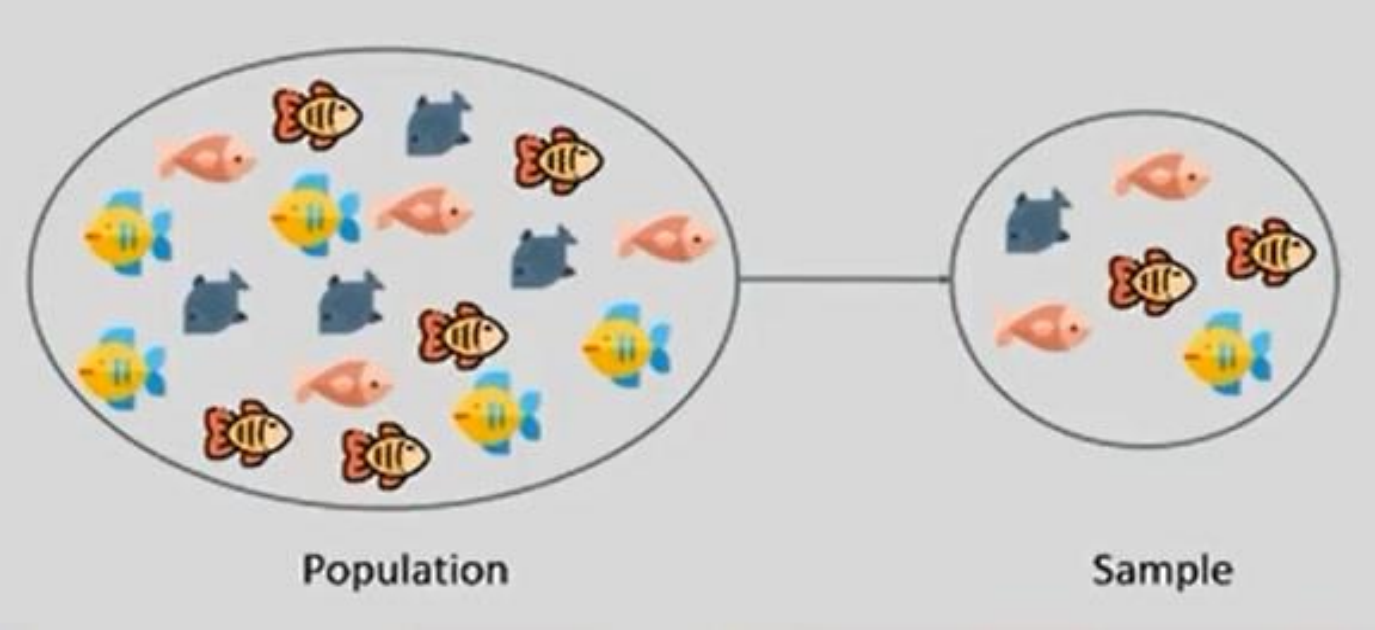
What is Statistics

The **latest sales data** have just come in, and your boss wants you to prepare a report for management on places **where the company could improve its business.** What should you look for? What should you not look for?

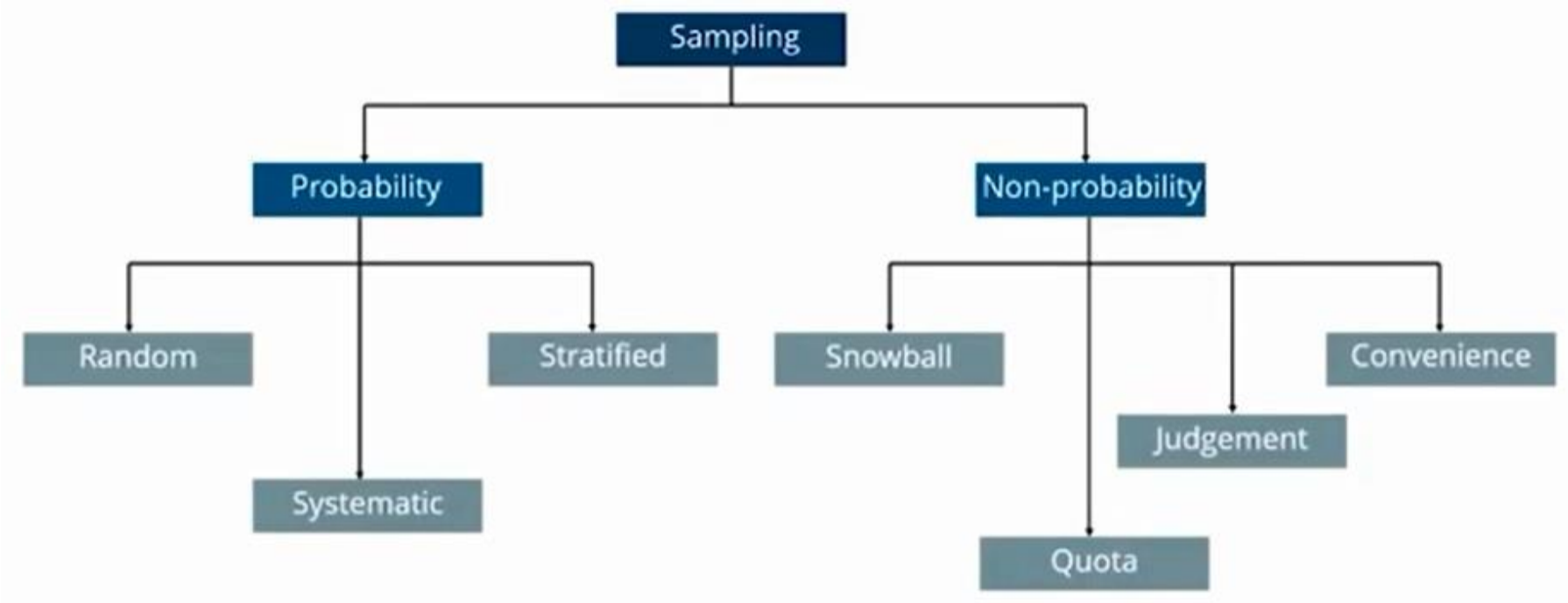
Basic Terminologies in Statistics

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

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



Sampling Techniques



Random Sampling

- **Random Sampling**
- **Systematic Sampling**
- **Stratified Sampling**

Each member of the population has equal chance of being selected in the sample.

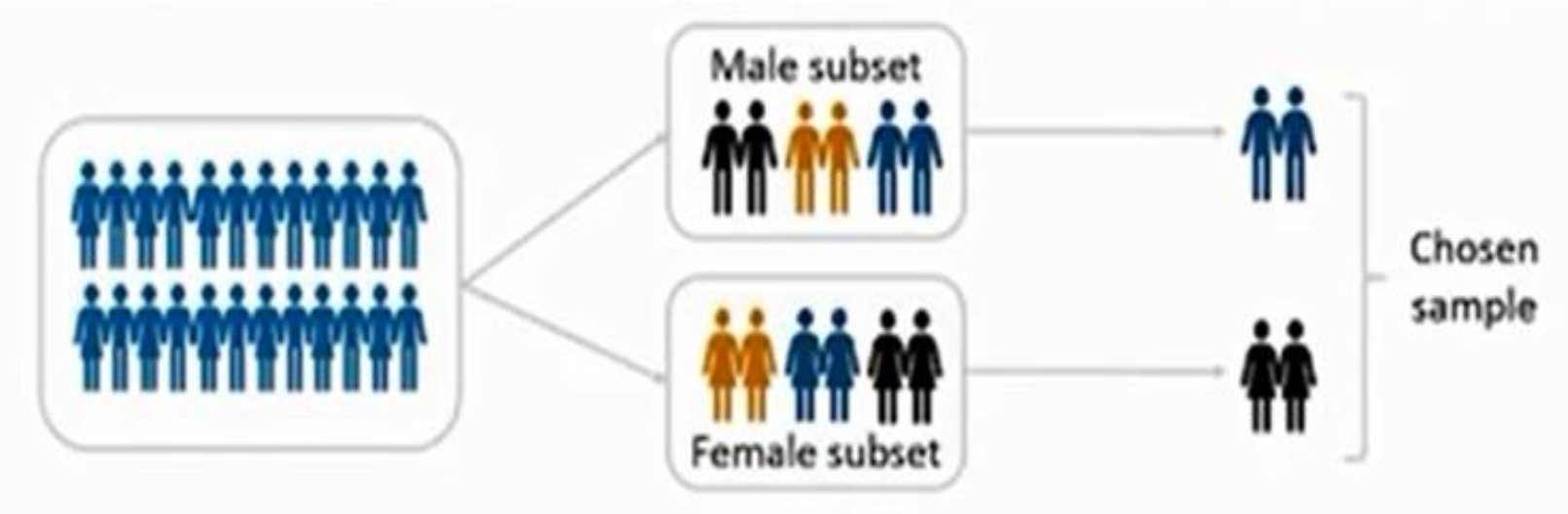
Systematic Sampling

❖ Systematic Sampling

- In **Systematic sampling** every **nth record is chosen** from the population to be part of the sample.

Stratified Sampling

❖ Stratified Sampling



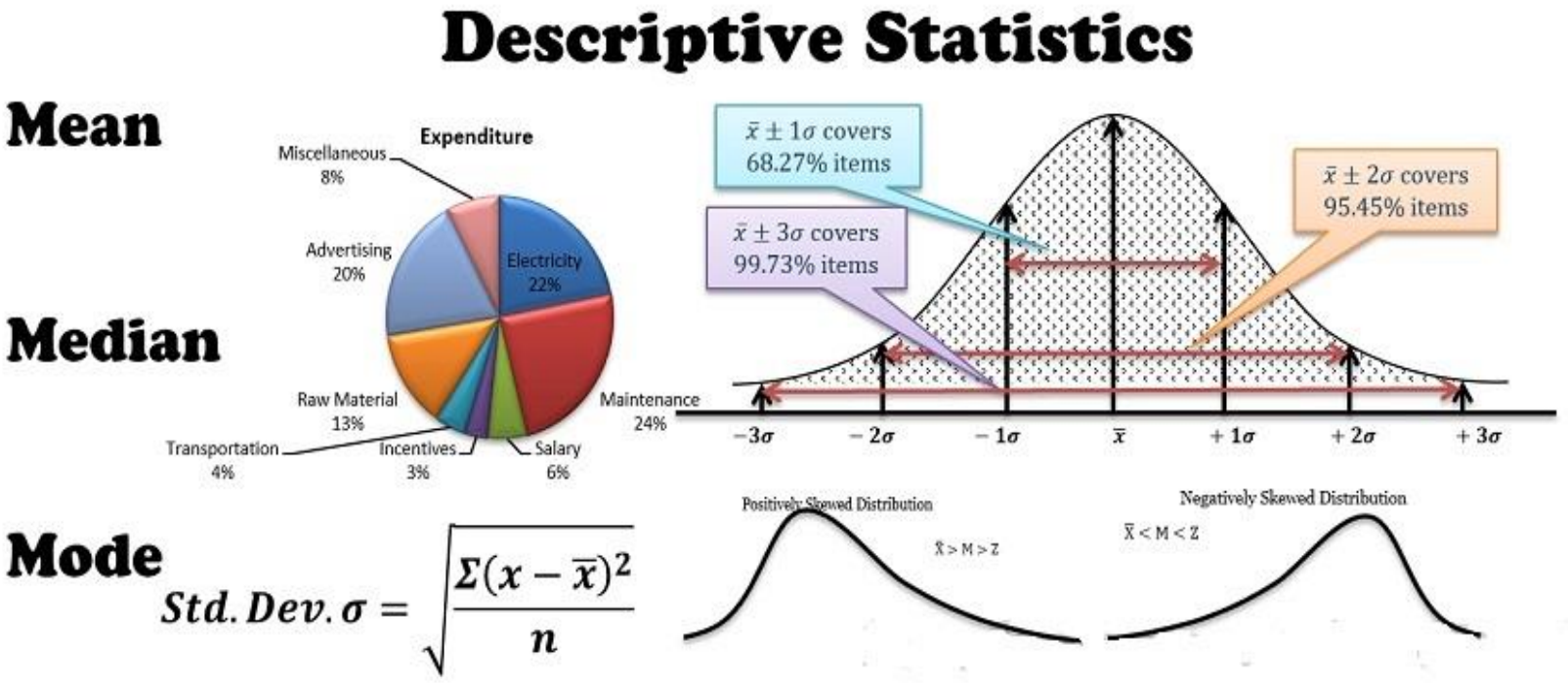
- **A Stratum** is a subset of the population that shares at least one common characteristic, in this case its **gender.**
- **Random Sampling** is used to select a sufficient number of subjects from each stratum

Different Types of Statistics

- **Descriptive Statistics**
- **Inferential Statistics**

Descriptive Statistics

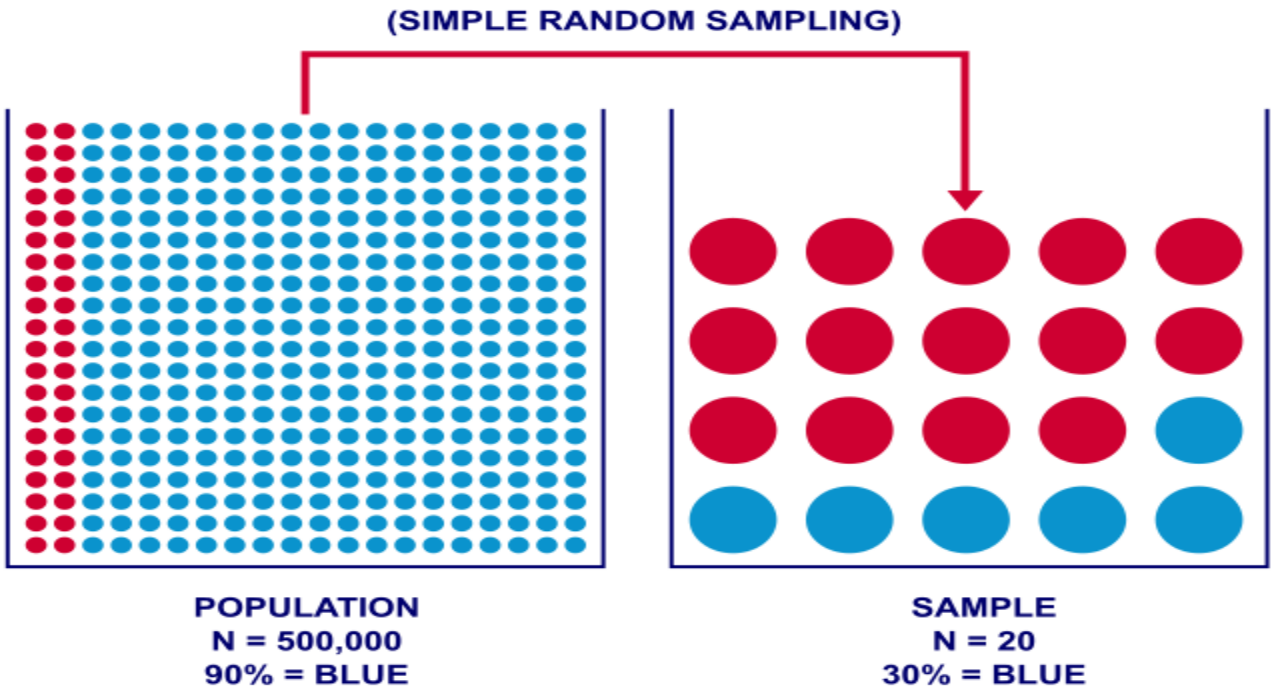
Descriptive Statistics uses the data to provide descriptions of the populations, either through numerical calculations, or graphs or tables.



Descriptive Statistics is mainly focused upon the main characteristics of data. It provides the graphical summary of the data.

Inferential Statistics

Inferential Statistics makes inferences and predictions about a population based on sample of data taken from the population in question.



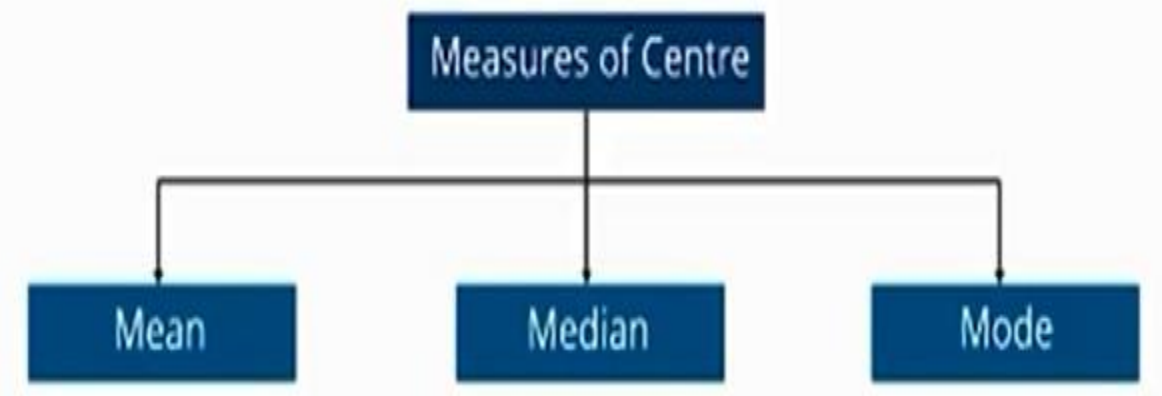
Inferential statistics, generalizes, a large dataset and **applies probability to draw conclusion.** It allows us to infer data parameters based on a statistical model using a sample data.

Descriptive Statistics

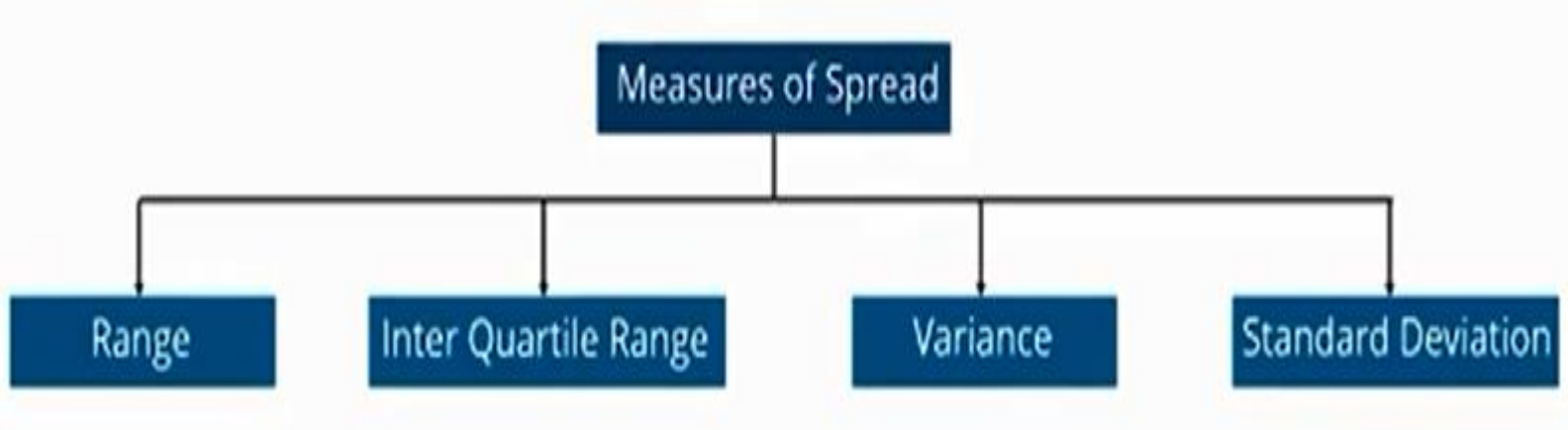
Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and the measures of the data.

Descriptive Statistics are broken down into two categories

1. Measures of Central tendency



2. Measures of Spread (Variability)



Mean

Measures of average of all the values in a sample is called **Mean**

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

To find the out the average horse power of the cars among the populations of cars, We will check and will calculate the average of all values.

(110+110+93+96+90+110+110+110)/8 =103.625

Median

Measure of the central value of the Sample set is called as **Median**

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

To find out the center value of the mpg among the population of the cars, arrange records in Ascending order, i.e. 21, 21, 21.3, 22.8, 23, 23, 23, 23
In case of even entries, take average of the two middle values i.e.
22.8+23/2=22.9

Mode

The value most recurrent in the sample set is known as **Mode**

Here is a sample dataset of cars containing the variables:

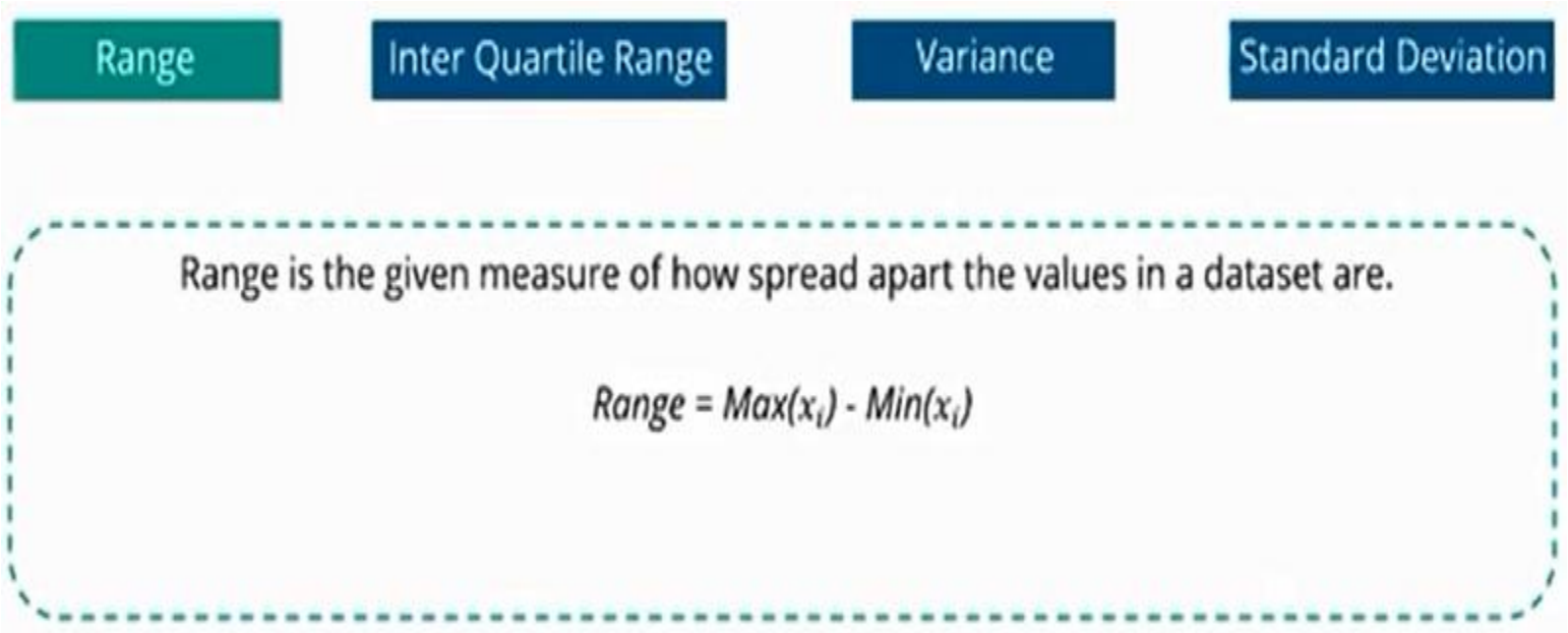
- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

To find out the most common type of Cylinder among the population of cars, check **the value which is repeated most number of time** i.e. Cylinder type 6

Measures of Spread-Range

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.



Measures of Spread-Quartiles

Range | Inter Quartile Range | Variance | Standard Deviation

Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.

Quartile Formula

$$\text{Lower Quartile (Q1)} = (N+1) \times \frac{1}{4}$$

$$\text{Middle Quartile (Q2)} = (N+1) \times \frac{2}{4}$$

$$\text{Upper Quartile (Q3)} = (N+1) \times \frac{3}{4}$$

Measures of Spread

How do you find the interquartile range?

We can find the interquartile range or IQR in four simple steps:

1. Order the data from least to greatest
2. Find the median
3. Calculate the median of both the lower and upper half of the data
4. The IQR is the difference between the upper and lower medians

Measures of Spread

Inter Quartile Range

Consider the marks of 100 students below, ordered from the lowest to the highest scores

Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

The first quartile (Q1) lies between the 25th and 26th.

$Q1 = (45 + 45) \div 2 = 45$

The second quartile (Q2) between the 50th and 51st.

$Q2 = (58 + 59) \div 2 = 58.5$

The third quartile (Q3) between the 75th and 76th.

$Q3 = (71 + 71) \div 2 = 71$

Measure of Spread

Inter Quartile Range

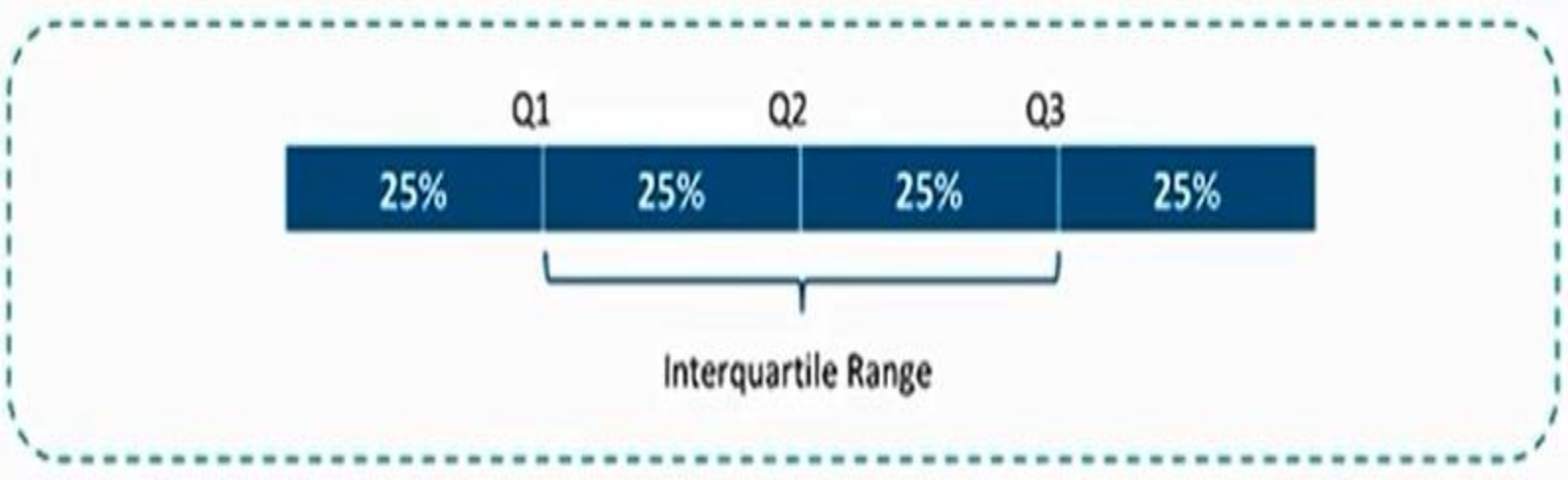
Inter Quartile Range (IQR) is the measure of variability, based on dividing a dataset into quartiles.

Quartiles divide a rank-ordered data set into four equal parts, denoted by Q1,Q2,and Q3, respectively.

The interquartile range is equal to Q3 minus Q1 i.e. IQR=Q3-Q1

Measure of Spread

Inter Quartile Range



Measure of Spread

Deviation

Deviation is the difference between each element from the mean.

Deviation = $(X_i - \mu)$

Measure of Spread

Standard Deviation

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

Measure of Spread

Standard Deviation– Use Case

Ross has 20 Dinosaur figures. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 1

Find out the mean for your sample set.

The Mean is:

$$\frac{9+2+5+4+12+7+8+11+9+3+7+4+12+5+4+10+9+6+9+4}{20}$$

$\therefore \mu = 7$

Measure of Spread

Standard Deviation– Use Case

Ross has 20 Dinosaur figures. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 2

Then for each number, subtract the Mean and square the result.

$(x_i - \mu)^2$

$(9-7)^2 = 2^2 = 4$
$(2-7)^2 = (-5)^2 = 25$
$(5-7)^2 = (-2)^2 = 4$
And so on...

$\therefore$ We get the following results:
4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9

Measure of Spread

Standard Deviation– Use Case

Ross has 20 Dinosaur figures. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 3

Then work out the mean of those squared differences.

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

$$\frac{4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9}{20}$$

$\therefore \sigma^2 = 8.9$

Measure of Spread

Standard Deviation– Use Case

Ross has 20 Dinosaur figures. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 4

Take square root of σ^2.

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

$\therefore \sigma = 2.983$


ThankYou
Merci
Salamat
Kop
Hvala
Tack
Xie
kun
Gra
Dankie
Kruthagnathalu
Arigatou
Dhonnobaad
Asante
Hain
Dhan
daa
Khopjai
Dhanyavad
Gomapsupnida
Grazzi
raibh
Gracias
Nandree
Blagodariya
Fyrir
Terima
Enkosi
Danke
dank
Euxaristo
Kun
Shukriya
Or
ederim
Falemínderit
Grazie
Go
Dhanyavaad
Spaas
Mul
Ačі
Shokriya
Ngiyabonga
Dziękuję
Todah
Mamnoon
Kasih
umesc
Sheun
Diolch
Kiitos
Shnorhakalutiun
Gamsahapnida
Dank
Takk
Waad
Daw
Dhanyavadaalu
Dakujem
krap
Te°ekkür
Dekuju/Dekujeme
Cam
or
al