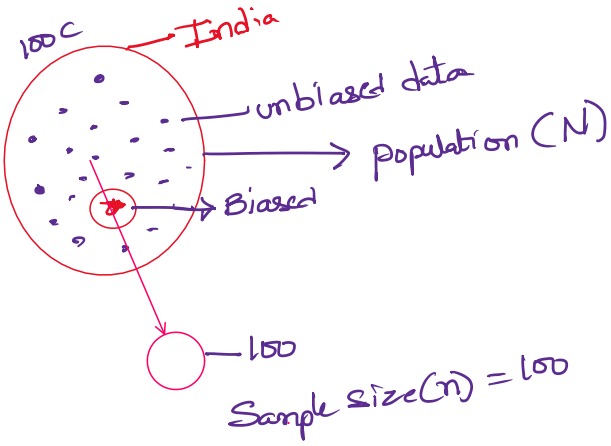
Basic Statistics

27 April 2024 10:44

Army: 1 week --> 5000 --Male --> No past or historical data

1. Indian
2. Age >= 21
3. Graduate:
4. Height ----> .....



6 feet ---> 5000 --> 3000
5 feet --> 5000 --> 500000

Currently, what is the average Indian height

Collect the data

The data which we cannot collect the entire portions --> Population data
We can collect only sample data from the entire population.

X -----> Randomly variable

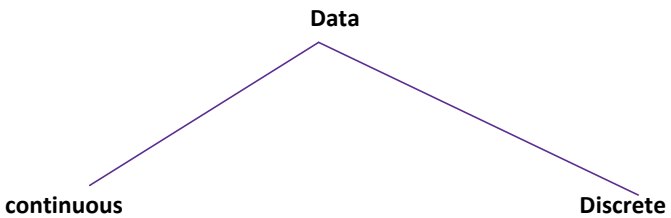
A	B	C	D
Height			
171		n	10
165		min	158
160		max	189
158		averag	170.3
172		median	170.5
189		mode	
168		std	
174		var	
176			
170			

From the **sample data**, whatever we calculates the statistical measures we will call it as "**Descriptive Statistics**"

From the sample data, using DS and applying some additional theory and estimating statistical measures on population data is called as "**Inferential Statistics**"

==========

What is **Descriptive Statistics?**



Any variable which is following a units

Any variable which is following a units

Ratio: Non Negative
Height, age, weight
Mass, volume, speed, distance
Time, mobile memory,

Interval: any values -inf to inf
Complexion, temperature, sensex

countable

# of calls
# of students
# of days
#

dassification/categorical

Nominal: Black, brown, gray,
Gender: M /F
exam: P/F, session: online/offline
designation:
True (2),False(1)
ordinal:

Excellent (5), average, Poor(1)

Unique data:
Adhar card, pan card, emp ID, Driving license, application number, passport, voter card, account number , phone number
Transaction number , order number,

The above data is helpful only to remove duplicates entries from the data.

==========

1. Central tendency:

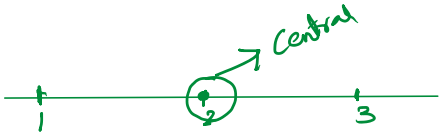
Mean:

$$\frac{\sum_{i=1}^{n}[x_i]}{n} = \frac{x_1 + x_2 + \dots x_n}{n}$$

Average ($\bar{x}$) → Sample

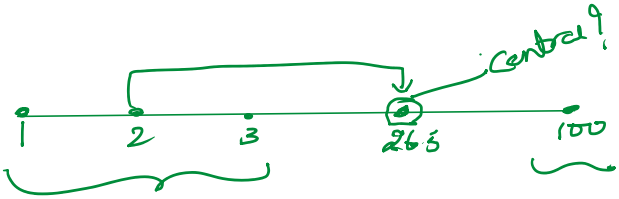
Mean — (μ) → Population

x: 1, 2, 3 $\quad \bar{x} = \frac{1+2+3}{3} = \frac{6}{3} = 2$



x: 1, 2, 3, 100 — outlier

$\bar{x} = \frac{1+2+3+100}{4} = 26.5$



Note: Mean will be always influences with outliers, in all the times mean may not be best centralized value

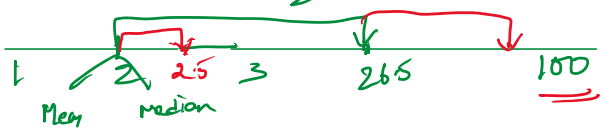
Median:

1. Sort the data in ascending order
2. If the n number of samples are even --> average of middle two values is our median
3. if the n is odd number --> middle value is our median

1 2 3 ⇒ $n=3$ → odd → Median = 2

1, 2, 3, 100 → $n=4$ → even → $\frac{2+3}{2} = [2.5]$ ✓

1, 2, 3, 500



Most of the times median is best central value than mean when we have the outliers

If we are able to identify the outliers , we can remove the outlier such that mean and median becomes same.
When we unable to identify the outliers, then median is the best central position.

Mode: Frequently repeated values/text will be considered as mode

H
gender
M
F
M
F
M
M
M
M
F
M
M
M
M

Mode = M → Bimode < M, F

Mode → numerical data

Mean, Median ≠ nominal data

==========

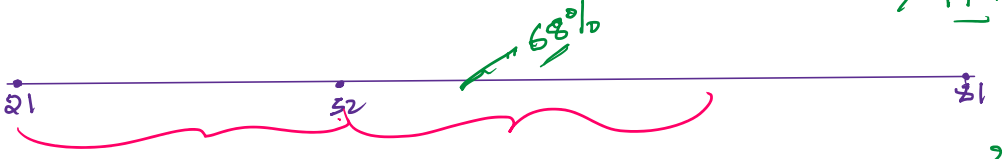
2. Spread / Dispersion

Range:
Max - Min = 81 - 21 = 60 → 21 − 81

Standard deviation: how much of deviation is existed from the centre value.

$$s = \sqrt{\frac{\Sigma [x_i - \bar{x}]^2}{n-1}} \rightarrow s^2 = \frac{\Sigma [x_i - \bar{x}]^2}{n-1}$$ → Variance

→ 99%. → Normal distribution



$s^2 = var$

$s = \sqrt{var}$

Most of the patients are in between 21 to 44 years old.

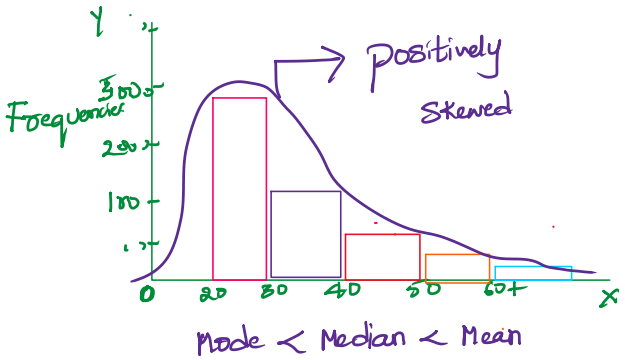
s = 11 , var = 121

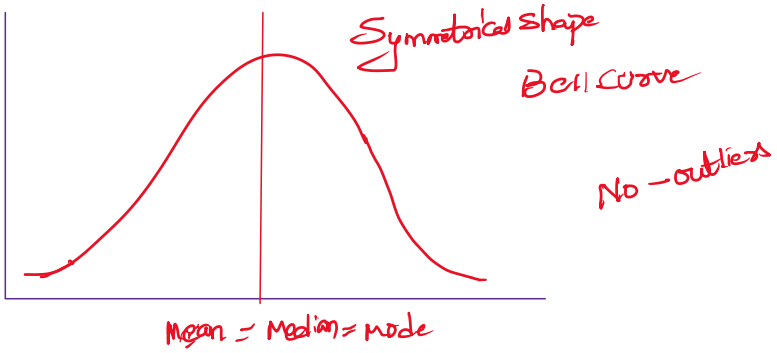
average	32.972		
sd	11.7413		
var	137.859	21.2307	44.7133
		-104.887	170.831 ✓

==========

3. **Shape of the data:**

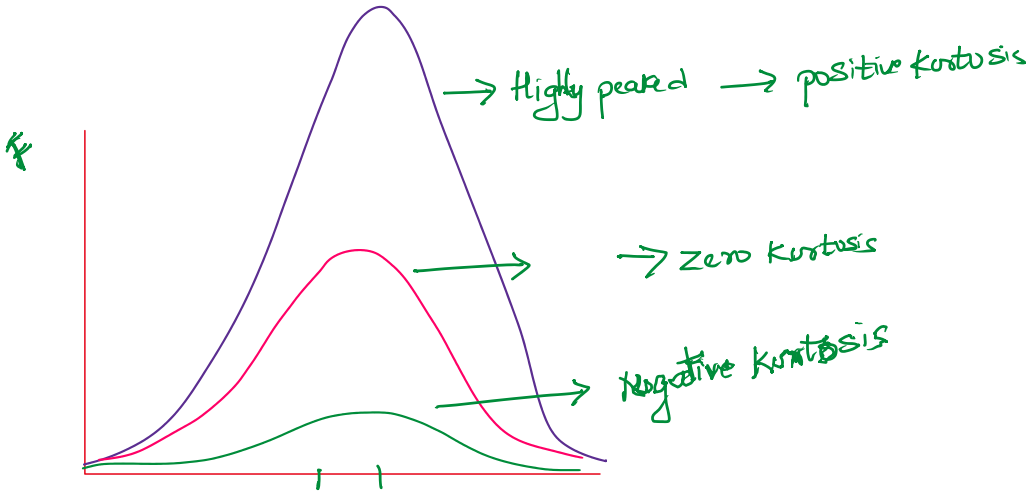
Histogram: it is graph which is visualized on X , Y axis, where x axis contains ranges or intervals of whole data and Y-axis contains its frequency.

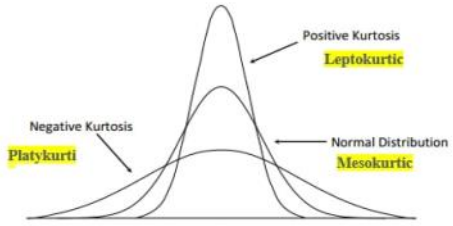


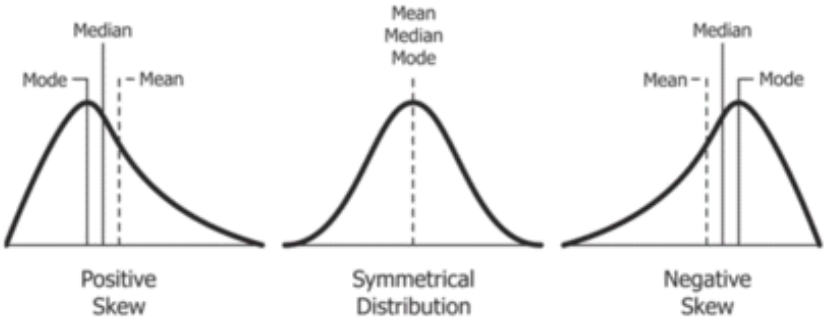


Skewedness values comes as > 0 --> positively skewed
Skewedness values comes as < 0 --> Negatively skewed
Skewedness values comes as = 0 --> Symmetrical shape

Kurtosis: It will calculates the peakedness of the data



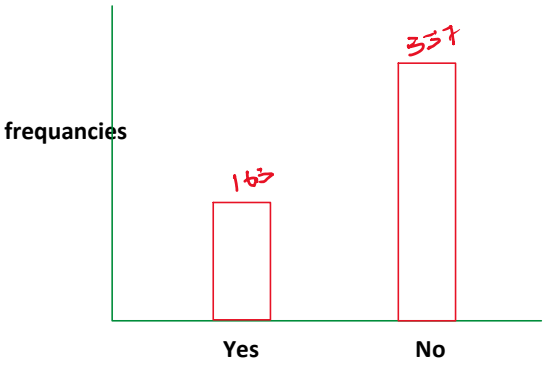




==========

Bar-graph:

It is for Discrete variables, on X - axis we will give all the categories and its frequencies on Y - axis
We are comparing between the categories to understand the which one is higher and which one is lower.



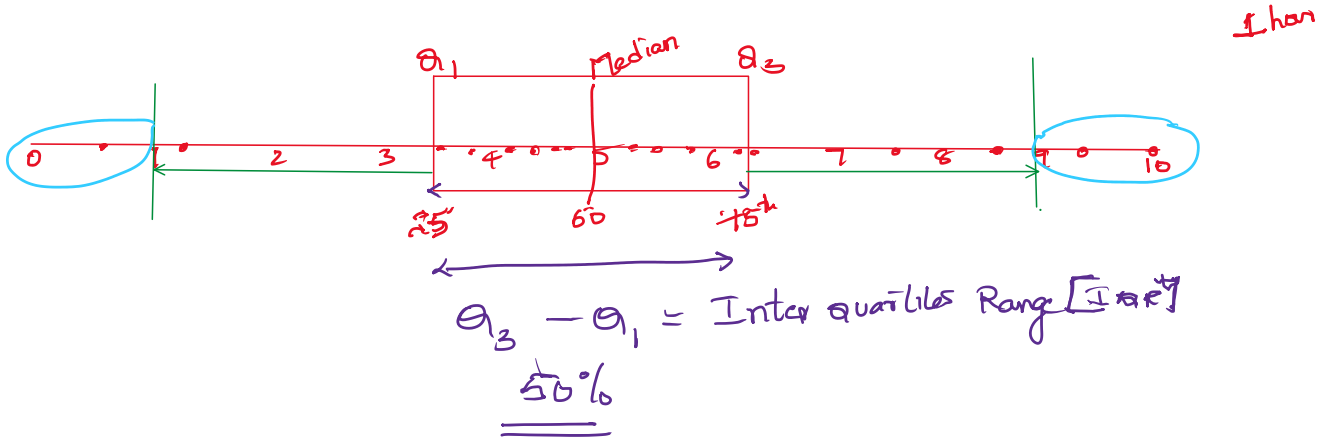
==========

Box plot: To Identify the outliers.

Outliers --> the values which are extremely far away from the remaining values.

What are percentage and percentiles?

	percentage--> individual inform			Ranks		percentiles --> 1 - 100th percentiles only				
600	1	580	0.966667	2						
	2	550	0.916667	4						
	3	560	0.933333	3						
	4	540	0.9	5						
	5	590	0.983333	1		100th percentile				
	6							75th percentile		Quartile 3
	7							50th percentile		Quartile 2
	8							25th percentile		Quartile 1
	9									
	200	420	0.7	200		1st percentile				



We will construct a scale on above of Q3 and below of Q --> whisker lengths

Upper whisker length = Q3 + (1.5 * IQR)

Lower whisker length = Q1 - (1.5 * IQR)

These borders will helps to decide that how many data points can be considered, the points which are above UWL and below LWL are treated as "OUTLIERS"

1	24000	2.5
2	25000	75
3	20000	78
4	30000	79
5	300000	80
average	79800	62.9

Box plots

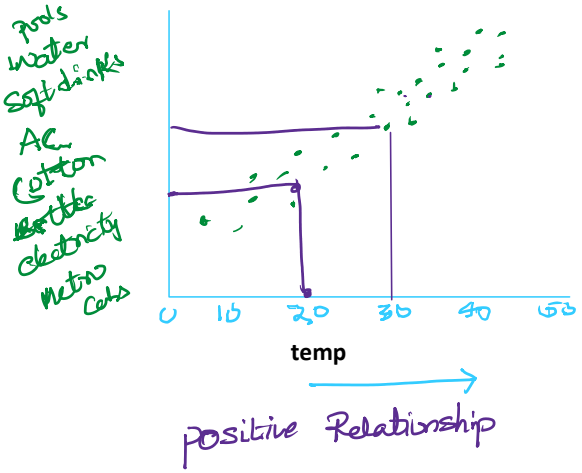
Note:

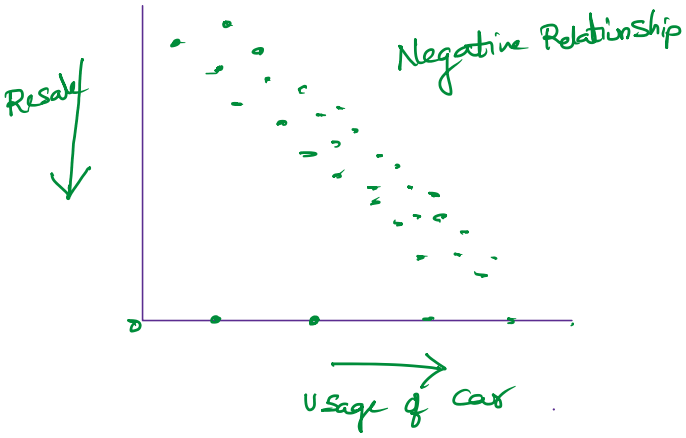
1. If sample size is very large and outliers are less, then we can remove the entire rows from the data.
2. If sample size is less and outliers are more, simple replace that value with median.

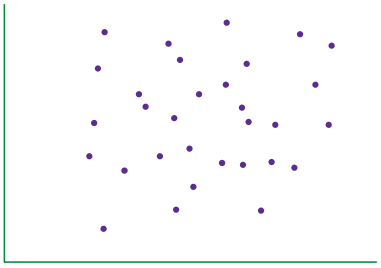
==========

Scatter plot:

It is a 2 dimension graph , where X axis and Y - axis both different variables
It will checks the relationship between two continuous variables





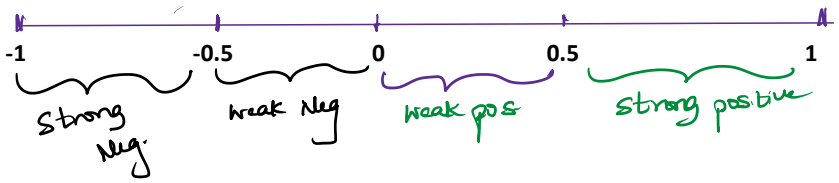


No – Relationship

We have statistical measure to decide how much of strong or weak in terms of positive / Negative

Correlation: lies in between -1 to + 1.

$$\gamma = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y}$$



==========