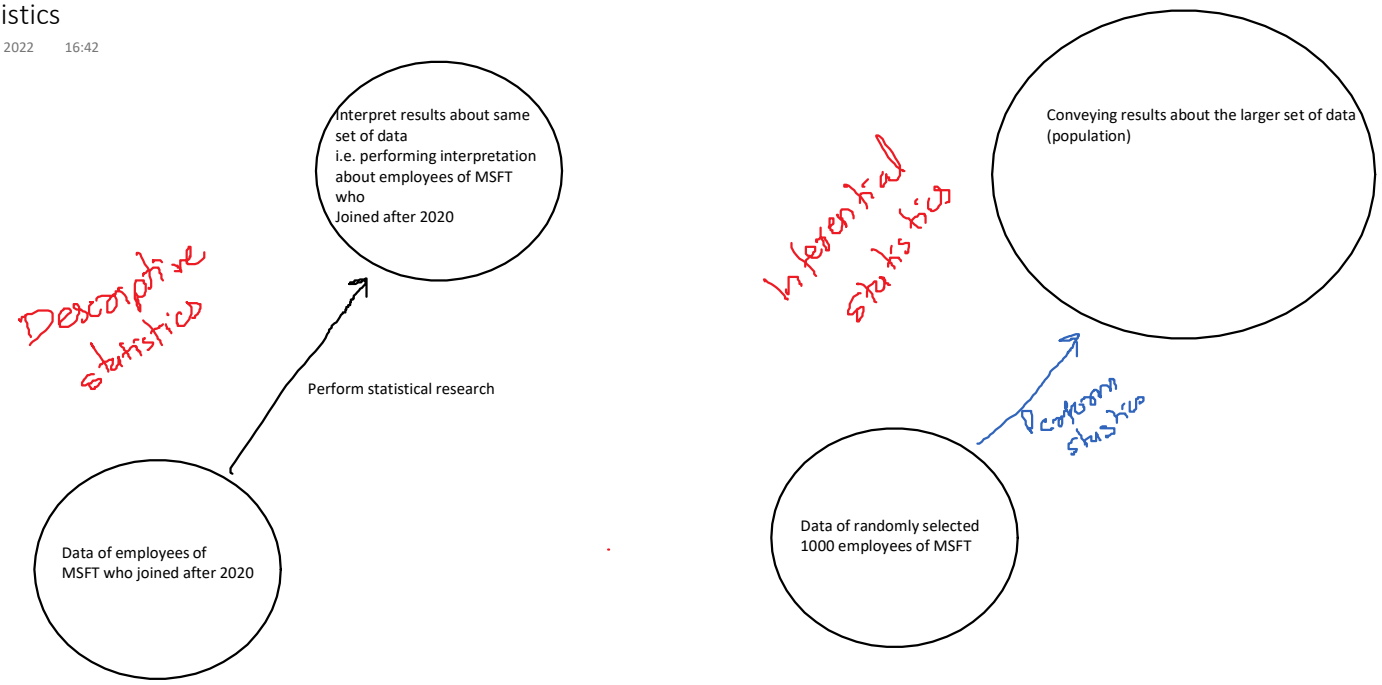
Statistics

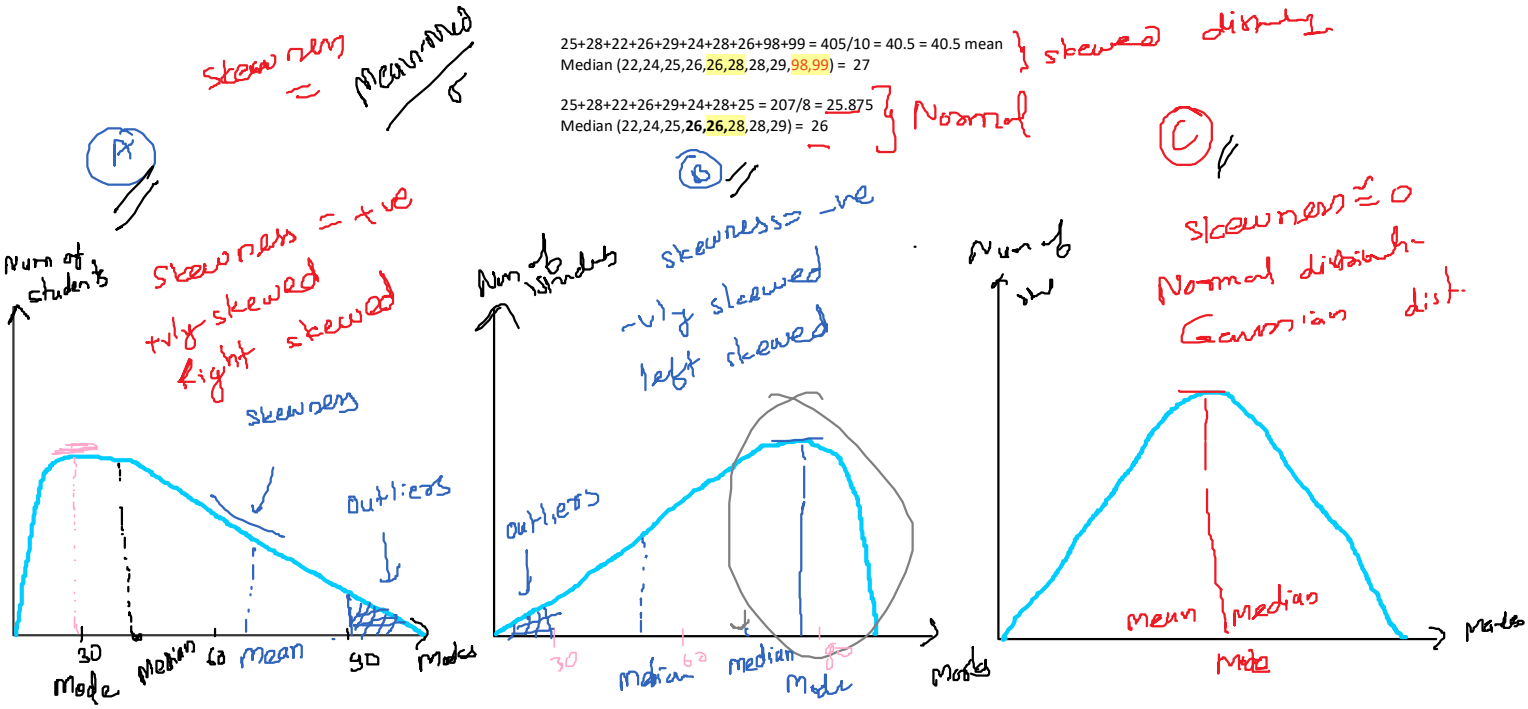
29 April 2022 16:42



Type of variable

- Numeric
  - Skewness - skewed = Median
  - Skewness - not skewed = Mean
- Categoric
  - Nominal = Mode
  - Ordinal = Median

Skewness



25+28+22+26+29+24+28+26+98+99 = 405/10 = 40.5 = 40.5 mean
Median (22,24,25,26,26,28,28,29,98,99) = 27
} skewed distribution

25+28+22+26+29+24+28+25 = 207/8 = 25.875
Median (22,24,25,**26,26,**28,28,29) = 26
} Normal

Measure of spread: variance, standard deviation and range

Delhi

18
22
23
26
29
34
36
37
42
46
43
42
38
37
34
32
28
26
25

$\bar{x} = 35°C$

Mumbai

22
22.2
23.2
23.6
24.2
22.8
23.3
23.6
23.8
23.5
24
22.9
23.1
23.2
23.6

$\bar{x} = 23.5°C$

Variance $= \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$

data = °C

$\bar{x} = 23.5\ °C$

$\sigma^2 = 9\ °C^2$

Machine	Temp	Vib	Mum	Pune	damaged
m1					Y
m2					N
m3					N
m4					N
⋮					⋮
m1000					N

$\sigma^2 = 0$

Standard deviation $= \sqrt{\text{variance}}$

$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$

A $1000

B $1000 $320

C $1000 $35

$\bar{x}$ $240

σ

Range → max − min

data → 25, 28, 37, 36, - - - , 78, 92, 96

Range = max − min

= 96 − 25 = 71

Percentage v/s Percentile

0th percentile

100th percentile

25,26,26,35,36,42,48,56,65,65,69,72,74,75,76,79,79,84,84,85,86

25% → 25th Percentile → Q1

50% → median

75th → Q3

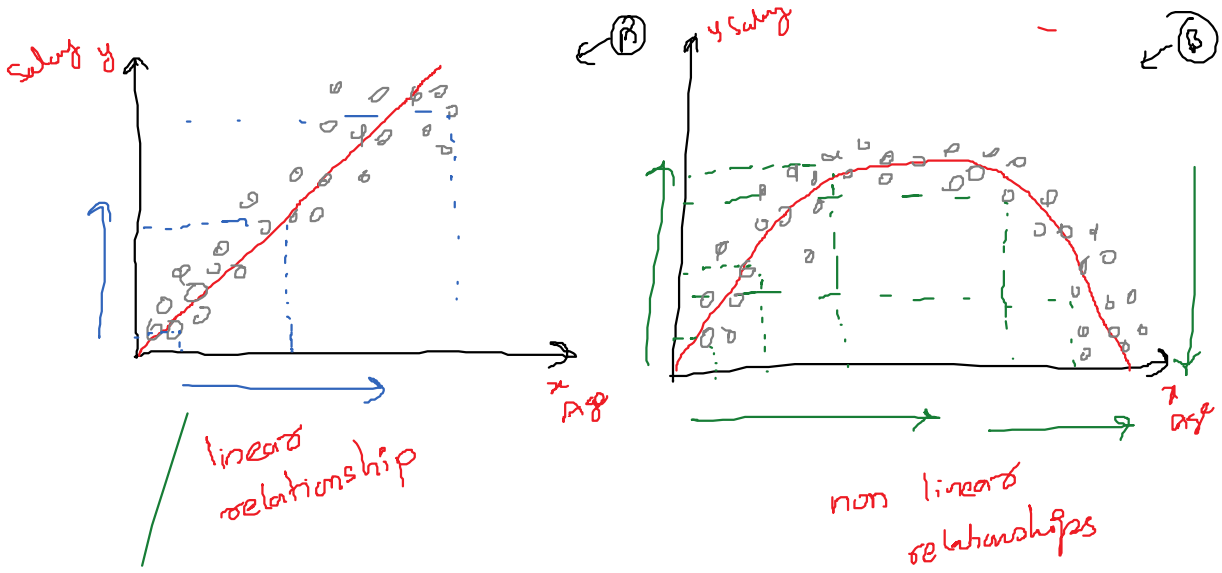
Percentage
1st student = (25*100)/100 = 25
Last student = (86*100)/100 = 86

Percentage
1st student = (25*100)/100 = 25
Last student = (86*100)/100 = 86

Percentile = num of values below *100 / total num of values

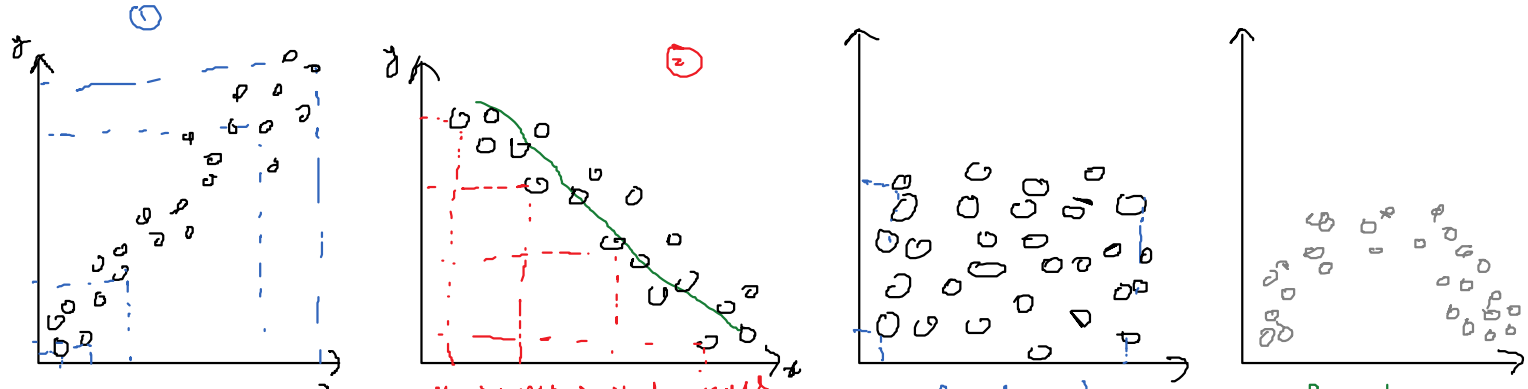
Marks = 20, percentile = 0*100/21 = 0
Marks = 25, percentile = 0*100/21 = 0
Marks = 86, percentile = 20*100/21 = 95.2381

25th Percentile 50'lo 75
Q1 median Q3
Q2



Correlation Analysis - correlation is a measure of linear relationship between two attributes (x and y)
- Value of correlation can range between -1 to +1

$$corr(x,y) = \frac{\frac{1}{n}\sum (x-\bar{x})(y-\bar{y})}{\sigma_x \cdot \sigma_y}$$



x increases → y increases
corr(x,y) = +ve
corr(x,y) ≥ +0.5
strong / Vgood
corr(x,y) is b/w +0.1 to +0.5
weak / slightly good

x increases → y decreases
corr(x,y) = -ve
corr(x,y) ≤ -0.5
strong / vgood
corr is b/w -0.1 to -0.5
weak / slightly good

x & y do not have any relation

x & y have non linear relation
bad correlation
corr ≅ 0 -0.1 to +0.1

A drug production company introduces a new drug ABC and based on testing/research they have done, they claim that this drug can cure a specific disease in 14 days. The business also claims that it cures 95% of patients in 14 days.

We need to verify their claim, we test this drug on 20 patients, below is the num of days it took for them to recover
Days = [12,14,12,15,18,12,10,11,13,15,14,12,15,14,17,15,16,15,16,14]

Objective: compare a sample mean with a population mean (a hypothetical value)

$$Z = \frac{x - M}{\sigma}$$

Z test - it is used to compare a population mean with a sample mean = requires pop std to be known
Null Hypothesis = sample mean is similar to population mean

Popmean = 14
CI = 95%, alpha = 0.05

$$t = \frac{x - \mu}{s/\sqrt{n}}$$

Population standard deviation (sigma) = not known

T test - it is used to compare a population with a sample mean - if population standard deviation is not known