Agenda

* Mean deviation
① Variance
② Std dev.

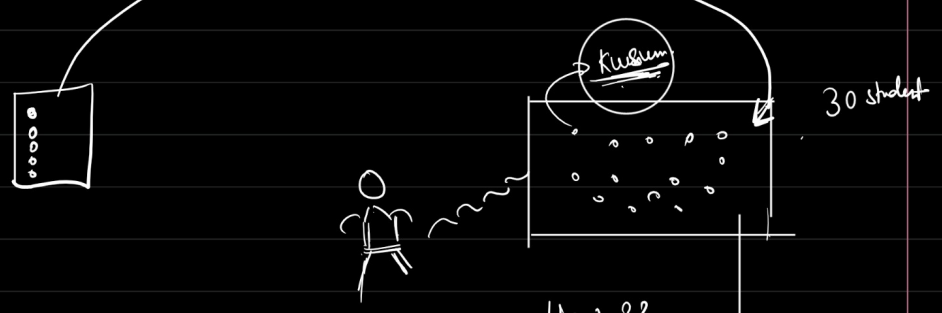
* Measures of Symmetricity
  * Skewness
  * Kurtosis

* Covariance and correlation.

* Probability.

inferential

Measures of spread → range, IQR, Percentile, Boxplot.



How ??

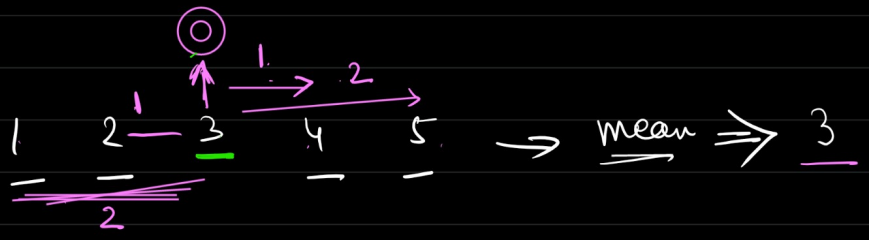
Compare Kusum's marks as compared avg of the class.

Avg marks = 21

* Mean deviation

each of the data point

mean



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

no. of dp

on an avg how much away each data point is away from mean value ??

$$\frac{1+2+1+2+0}{5} = \frac{6}{5} = 1.2$$

* Variance → the average of the squared differences from the mean

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

(μ → population mean)

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

$N \rightarrow n$

$\mu \rightarrow \bar{x}$

$N \rightarrow n-1$ (denominator)

Variance

→ Calculate mean

→ for each no in data, Subtract the mean and no.

→ Square the difference

→ Calculate the avg square of difference

data = { 1, 2, 3, 3, 4, 4 }

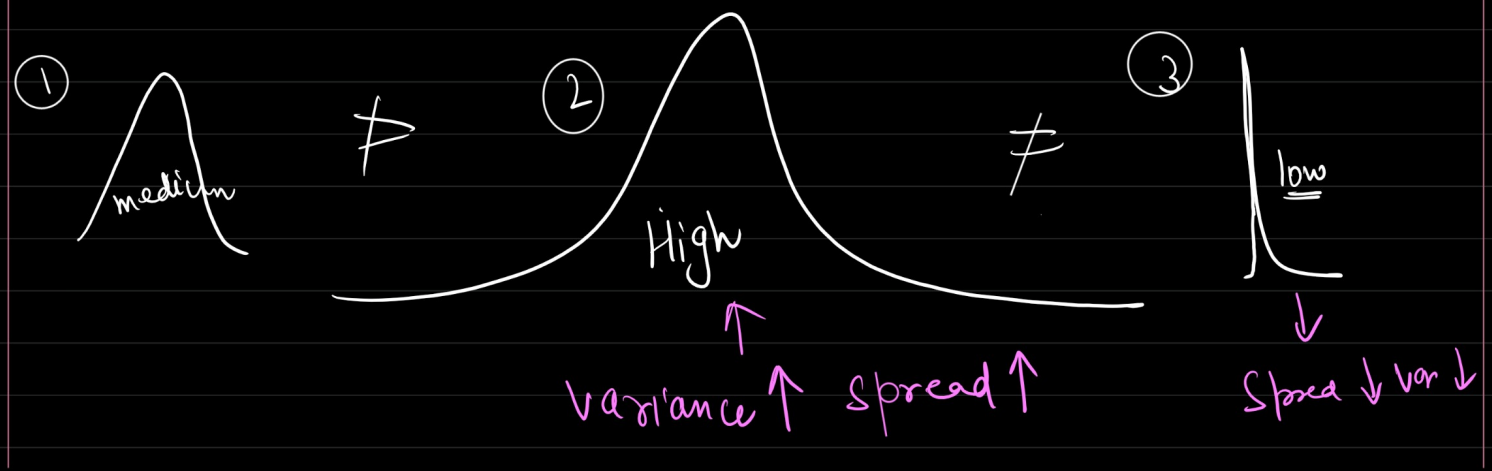
(Part of some sample)

x	$\bar{x}$	$x - \bar{x}$	$(x-\bar{x})^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.68
3	2.83	0.17	0.03
3	2.83	0.17	0.03
4	2.83	1.17	1.37
4	2.83	1.17	1.37
			6.82

$\frac{\sum x_i}{n} \Rightarrow \frac{17}{6} \Rightarrow 2.83$

$$Var_{sample} = \frac{6.82}{n-1} = \frac{6.82}{5}$$

⇒ 1.37



* Standard deviation

↳ Standard dev^n is a measure of how spread out numbers are.

↳ Square root of Variance.

$$\text{Std} = \sqrt{\text{var}} = \sqrt{1.37} = 1.17$$

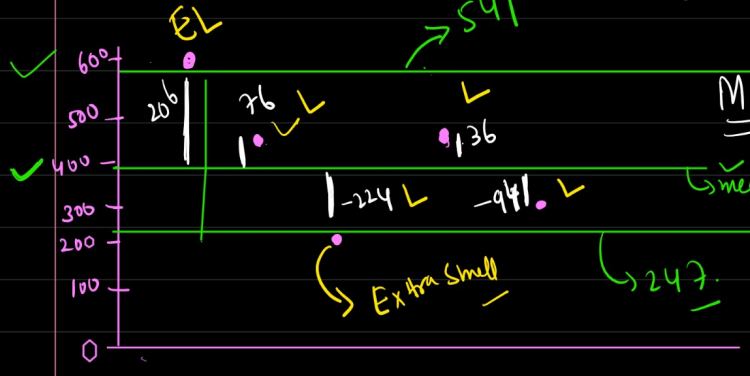
Standard devn of population $\Rightarrow \sigma = \sqrt{\text{var}_p}$

" " " " $\Rightarrow s = \sqrt{\text{var}_s}$

Use Cases

This is the class

Height → 600cm (Soumya), 470cm (Rahul), 170cm (Deep), 430cm (Farhan), 300cm (Sumana)



$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

$$= 394$$

↳ mean

$$\sigma^2 = 206^2 + 76^2 + (-224)^2 + (36)^2 + (-94)^2$$

$$\sigma^2 = 21704$$

Variance → 21704

Standard devn is a standard way of knowing how normal, large, extra large is something.

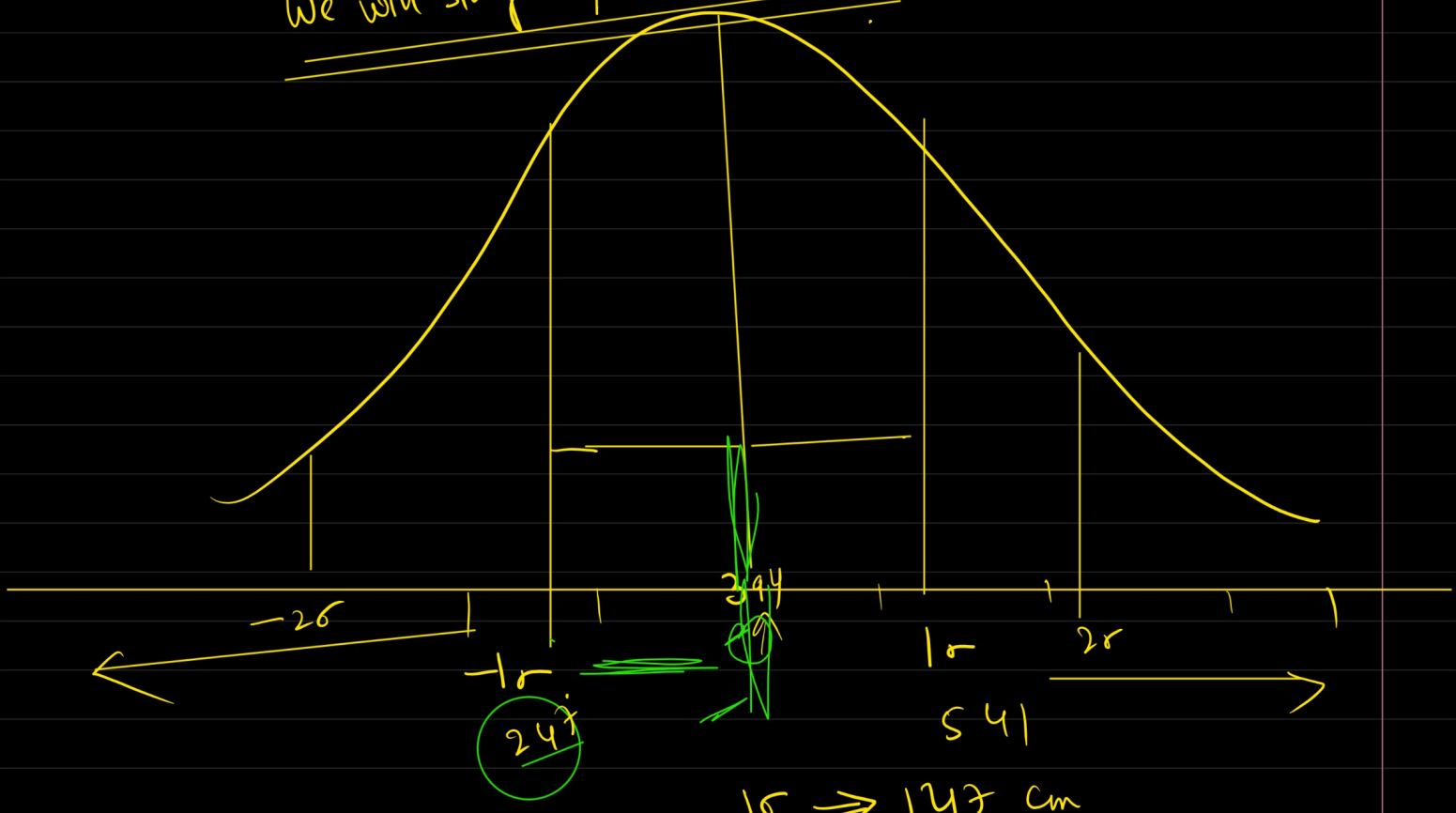
↓

Standard devn → $\sigma = \sqrt{21704} = 147$

$394 + 147 = 541$

$394 - 147 = 247$

We will study again in N.D



1σ ⇒ 147 cm

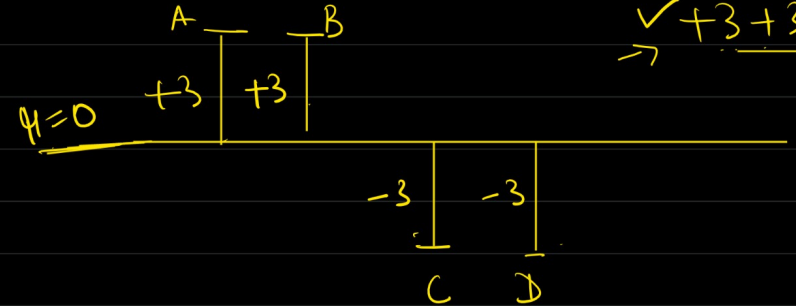
What was the mean deviation

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

To calculate standar dev.

$$\sigma = \sqrt{\text{variance}}$$

$(x-\mu)^2$ → Why.

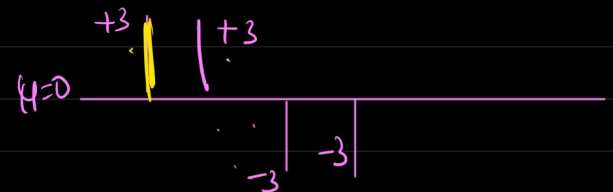


$$\frac{+3+3-3-3}{4} = 0$$

+ve and -ve negating

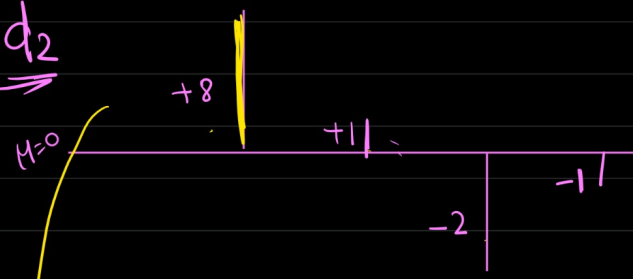
Can I take abs → $\frac{|3|+|3|+|-3|+|-3|}{4} = \frac{12}{4} = 3$

d_1 =



mean devn ⇒ 3 ✓ Mean devn

d_2 ⇒



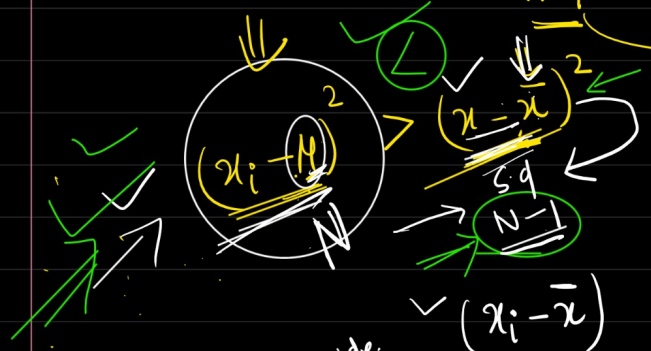
$$\text{mean devn} = \frac{|+8| + |+1| + |-2| + |+1|}{4} = 3$$

$$\sqrt{\frac{3^2 + 3^2 + (-3)^2 + (-3)^2}{4}} = \sqrt{\frac{36}{4}} = 3$$

$$\sqrt{\frac{8^2 + 1^2 + (-2)^2 + (-1)^2}{4}} = \frac{\sqrt{64+1+4+1}}{4} = 4.4$$

* $\text{Var}_{\text{sample}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$ Why $n-1$?? $\frac{(x_i - \mu)^2}{N}$

→ Bessel correction / unbiased estimator



We use $n-1$ rather than n is because sample variance will be unbiased estimator.

↳ $x_i - \mu$ ✗ (you don't have access to pop)

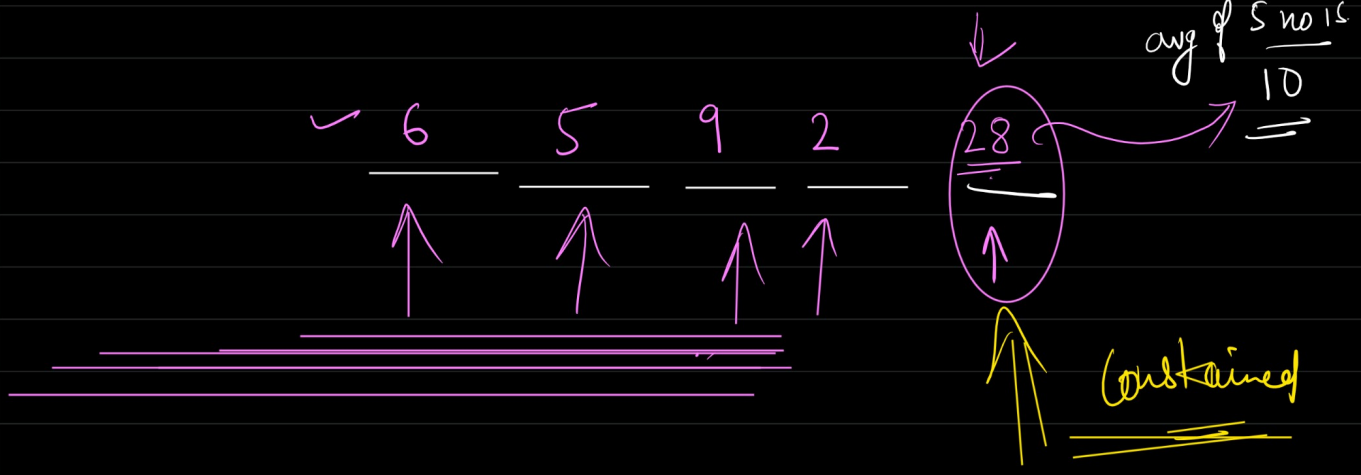
↳ Why − N−1 only

Why not N-2, N-3, N-4, N-5

degree of freedom

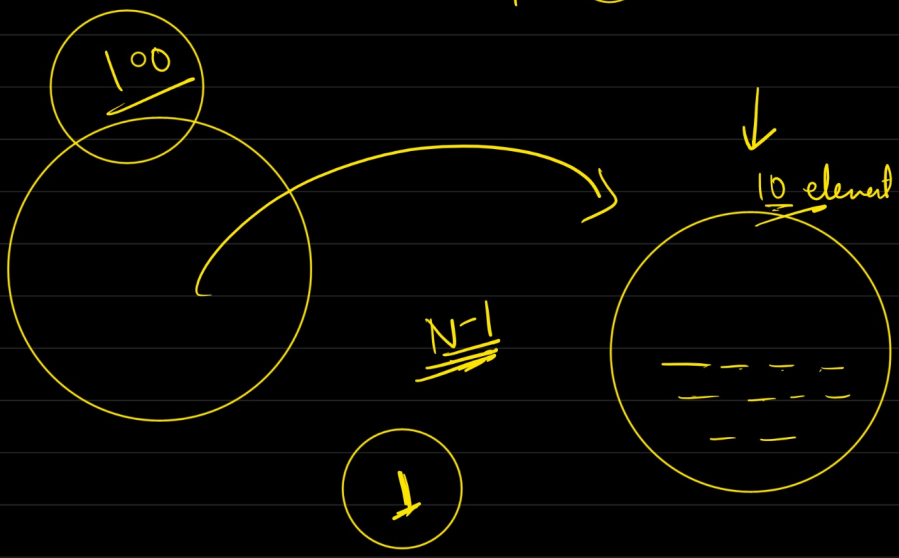
↓

maximum no of logical independent variables.



N N-1

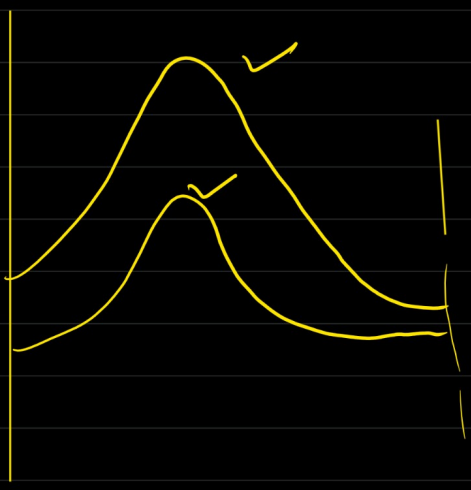
avg 20

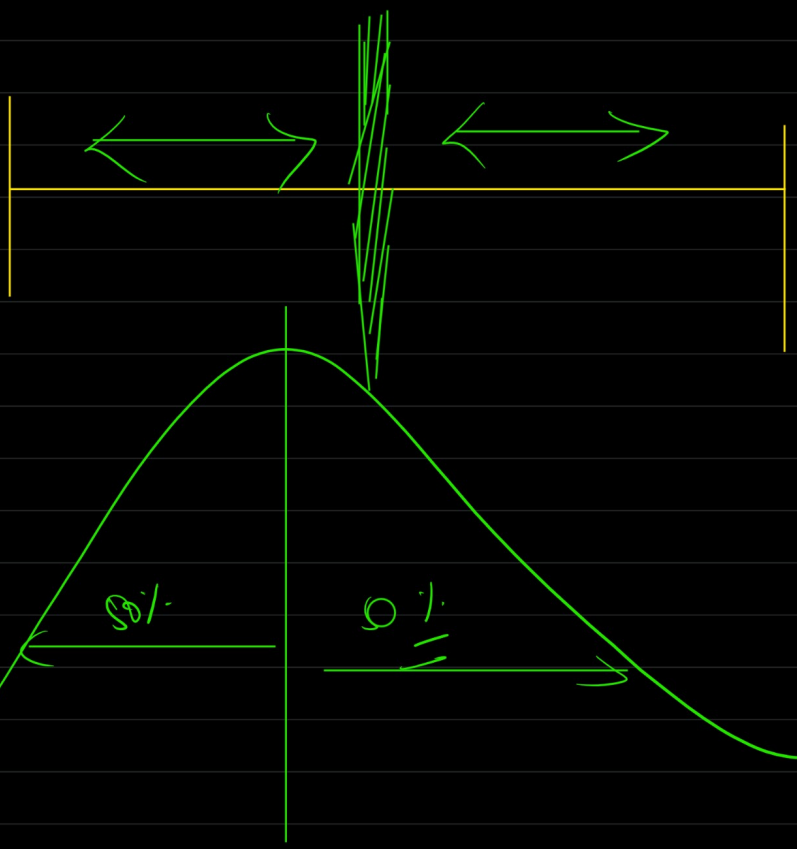


No. of positions are changeable

* Measures of Symmetricity

Descriptive Stats
- MCT
- MD
- Measures of Symmetricity



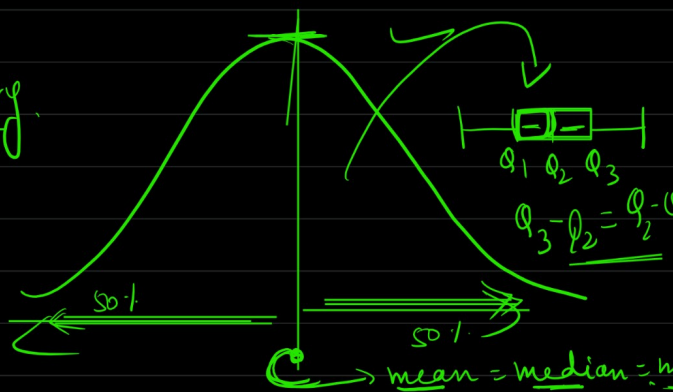


① Skewness

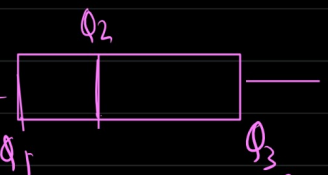
② Kurtosis

* Skewness → dataset's symmetry.

Skewness = 0



Skewed | non-Symmetric

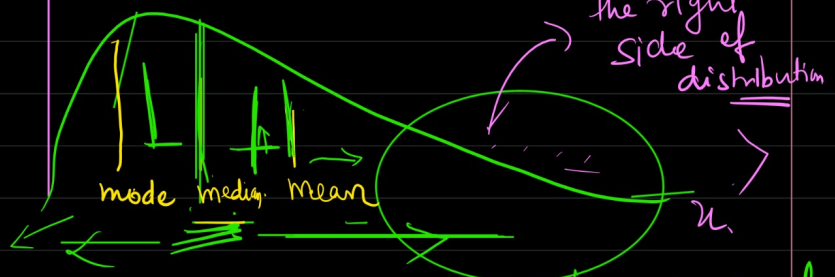


① Positive Skewed (Right Skewed data)

$Q_3 - Q_2 \geq Q_2 - Q_1$

tail is on the right side of distribution

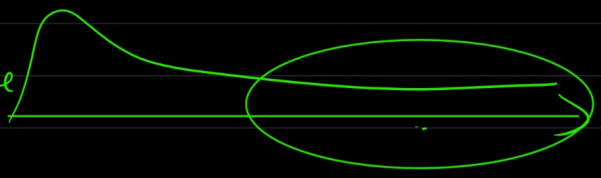
median → physical midpoint



mean ≥ median ≥ mode

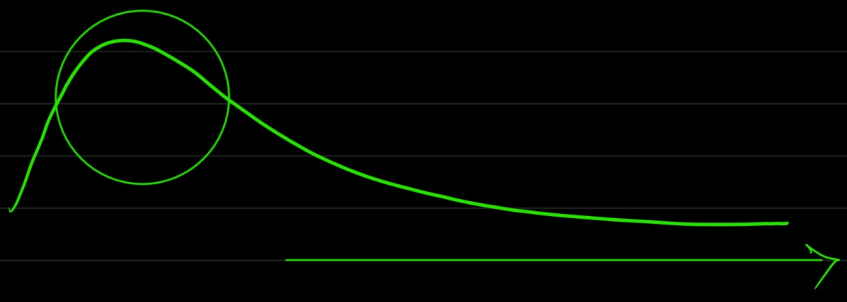
median — 50% | 50%

data is incline

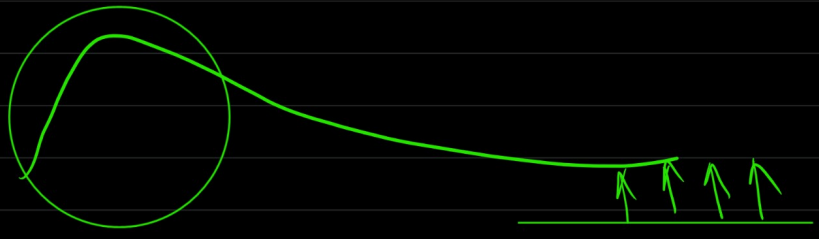


mean →

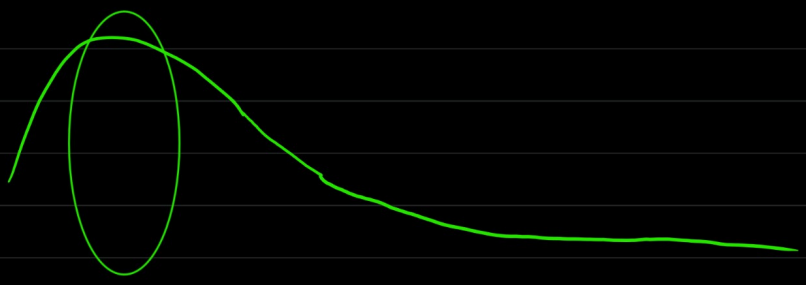
Example → Distribution of salary in a company.



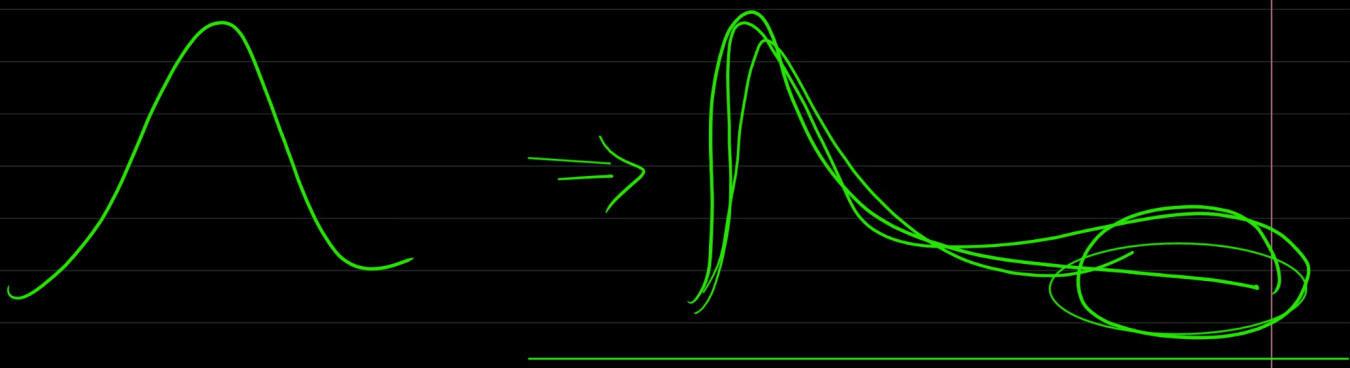
Distribution of marks in difficult exam



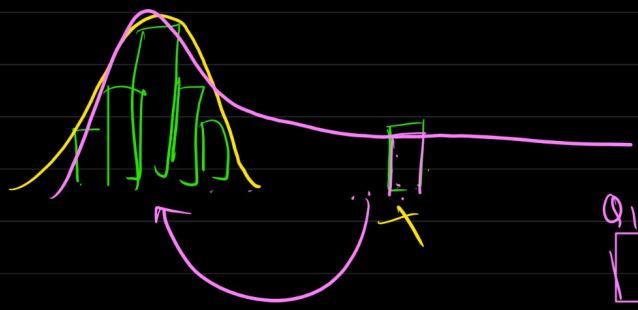
Article 370



Why ?

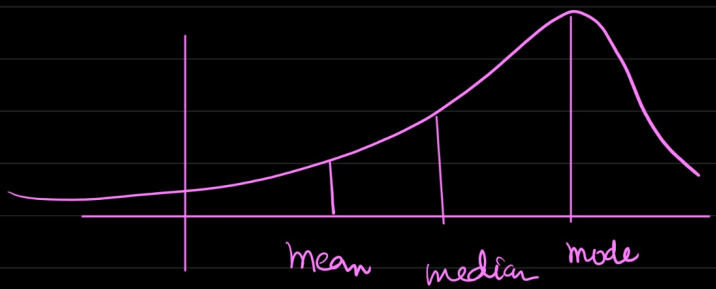


Outliers



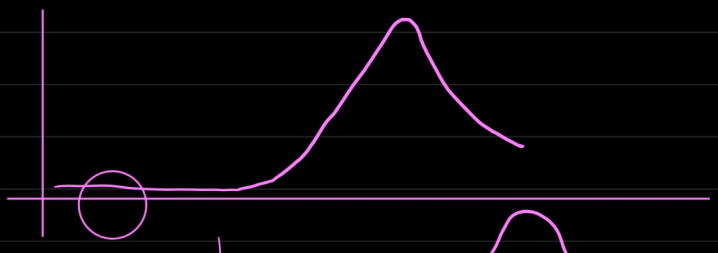
$Q_2 - Q_1 \geq Q_3 - Q_2$

② Left skewed data

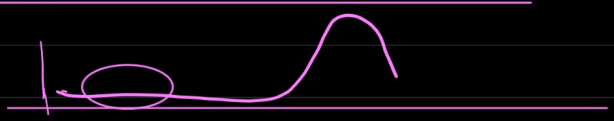


mode $\geq$ median $\geq$ mean

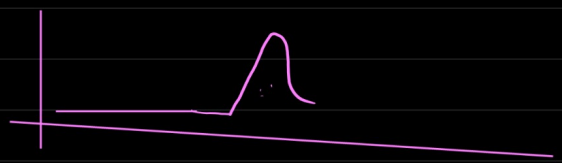
① Death rate

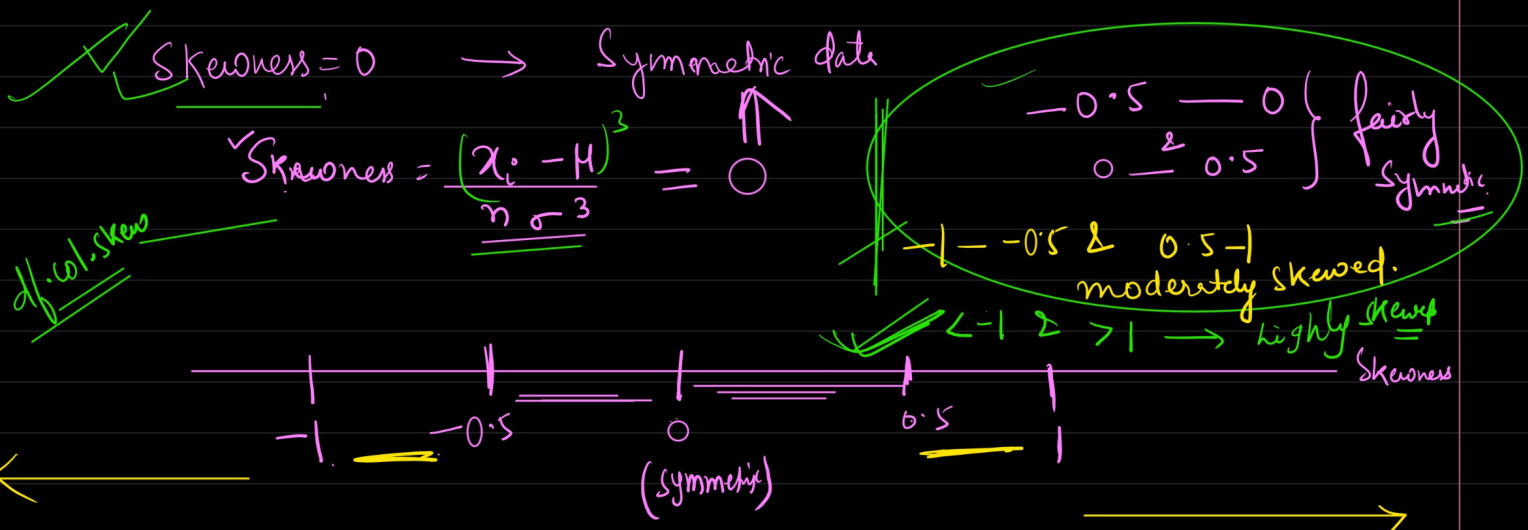


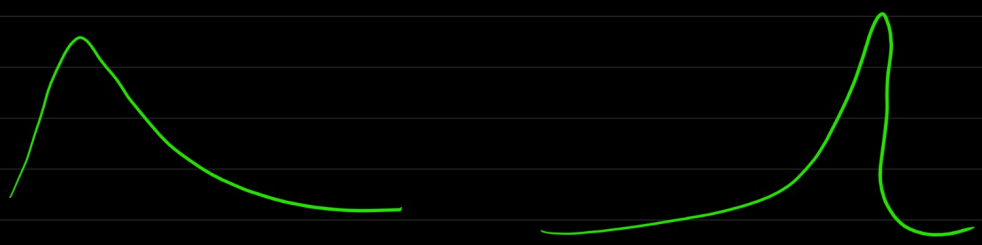
② Marks in easy exam



③ Wealth distrib







Reasons:-

→ Many ML models wants data to be symmetric

→ Visualization → Dist plot
→ Q-Q plot.

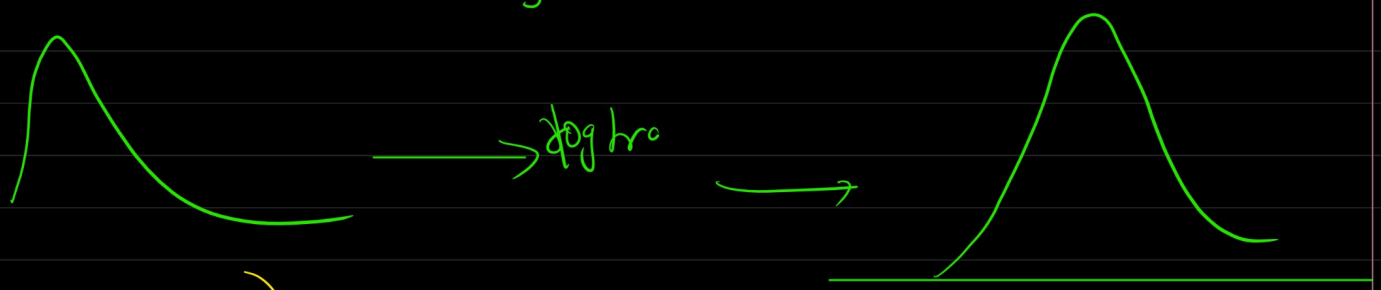
Statistical → Skewness df.col.skew

< -1 $> +1$

Transformation

- treat the outlier
- Box-Cox transformation | Yeo Johnson
- Exponential transf

↳ reciprocal
↳ log transformation



(1st moment)

$$\text{mean} = \frac{\sum (x_i - 0)^1}{N}$$ → mean is computed taking 0 as reference

(2nd moment)

$$\text{Var}_p = \frac{\sum (x_i - \mu)^2}{N}$$ → Here mean is the reference

(3rd moment)

$$\text{Skewness } q_3 = \frac{\sum (x_i - \mu)^3}{N \sigma^3}$$ → mean is the reference

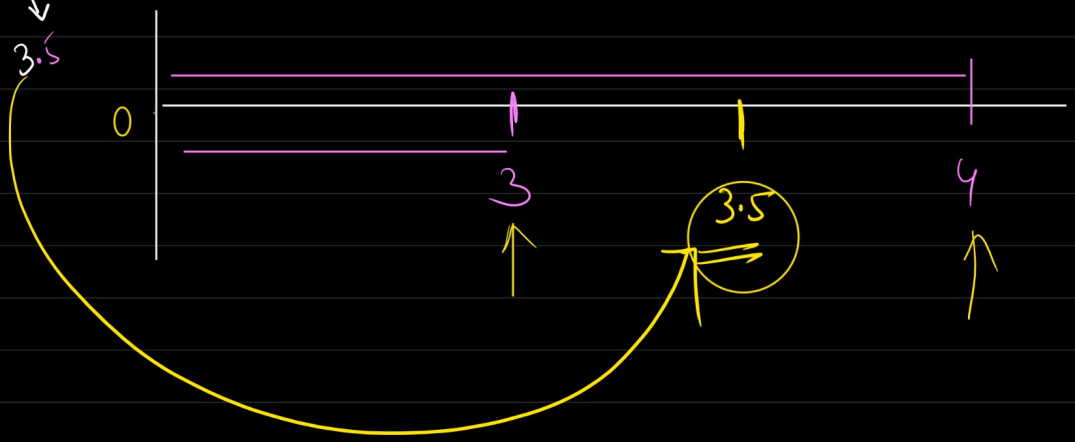
4th moment

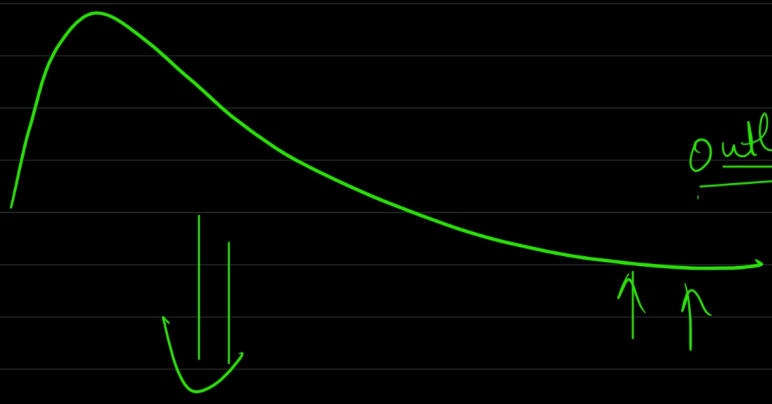
$$\text{Kurtosis } q_4 = \frac{\sum (x_i - \mu)^4}{N \sigma^4}$$

N
x_i
$\sum$

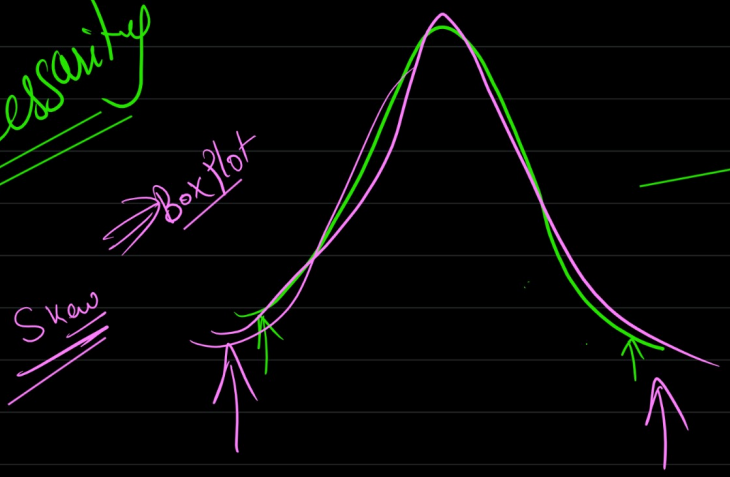
mean $\frac{\sum (x_i - 0)}{N}$ → On an avg how much a specific value or data point lie away from origin

3 & 4
↓
3.5





Not necessarily

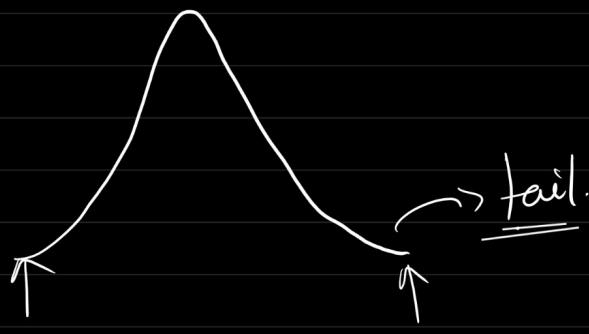


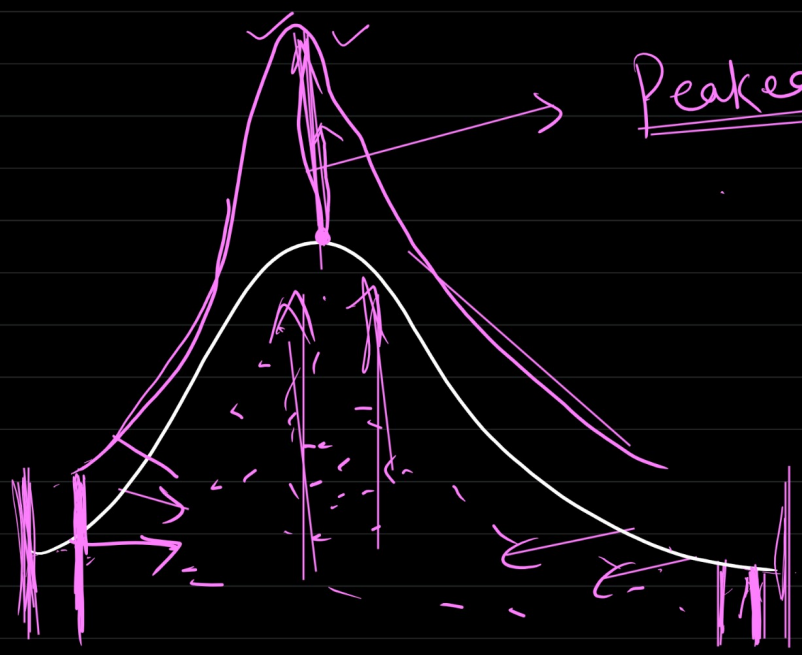
It can not have outliers ??

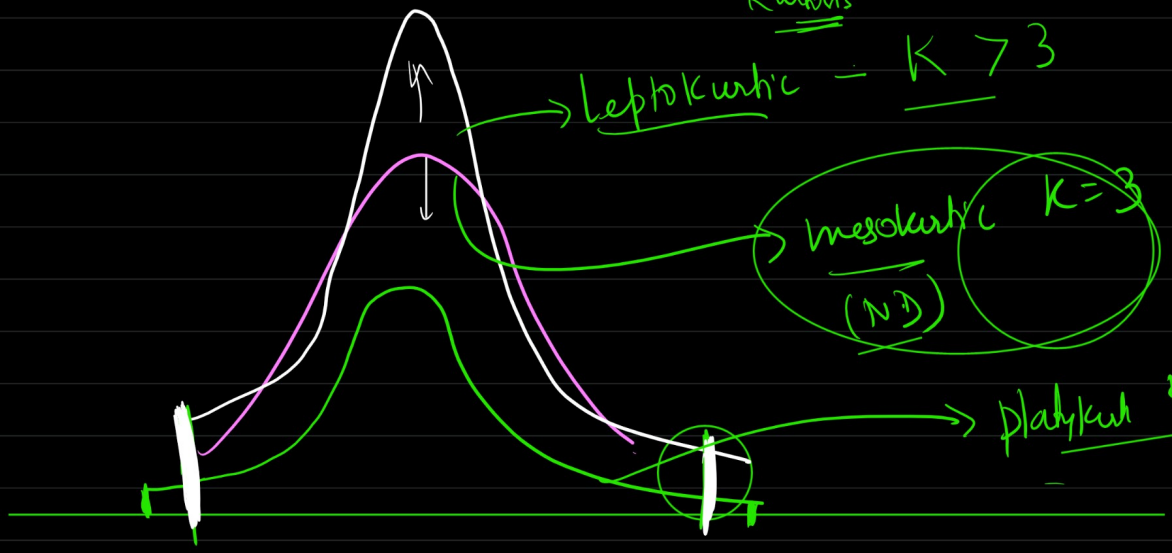
* Kurtosis

$$q_4 = \frac{(x_i - \mu)^4}{N s^4}$$

→ tail of the distribution | fatness of the tail.







✓ $K > 3$ → Leptokurtic → tail is fat → Many outliers

$K = 3$ → Meso

✓ $K < 3$ → Platykurtic