* Agenda

- → Quartiles
- → IQR
- → Boxplot
- → Variance
- → Std dev.
- * Measure of Symmetricity
- * Co-relation & Covariance
- * Prob dist.
- * CLT

* Quartile -

Problem with range → it is affected by outliers.

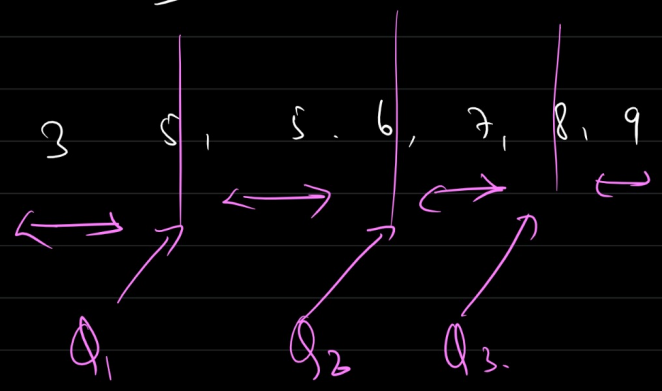
* Quartiles are the values that divides a list of Nos into quarters

→ Put the no in order
→ Then cut the no into 4 equal parts
→ The quartiles are at the cut.

Eg: 6, 8, 5, 5, 7, 3, 9.

Order - 3, 5, 5, 6, 7, 8, 9

Cut the no. into quarter



$Q_1 - 5$
$Q_2 - 6$
$Q_3 - 8$

Eg: 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4

total_no = 11 ↑ Q_1 ; ↑ Q_2 ; ↑ Q_3

if total no odd

Q_1 - $\frac{n+1}{4}^{th} = \frac{11+1}{4} = \frac{12}{4} = 3rd$ no

$Q_1 : 1$

$Q_3 = \frac{3(n+1)}{4}^{th} = \frac{3(11+1)}{4}^{th} = \frac{3\times12}{4} = 9^{th}$ ob.

$= 3.$

$Q_2 = \frac{(n+1)}{2}^{th} = \frac{11+1}{2}^{th} = 6^{th}$ no. $= 2$

if no is even

$Q_1 = \frac{n}{4}^{th}$ no

$Q_3 = \frac{3n}{4}^{th}$ no

$Q_2 = \frac{\frac{n}{2}^{th} + \left(\frac{n}{2}^{th}+1\right)^{th}}{2}$

* $Q_2 \Rightarrow$ Median

* Five Point Summary

Q_0 (min) — 0^{th} percentile

Q_1 — 25^{th} percentile

Q_2 — 50^{th} percentile

Q_3 — 75^{th} percentile

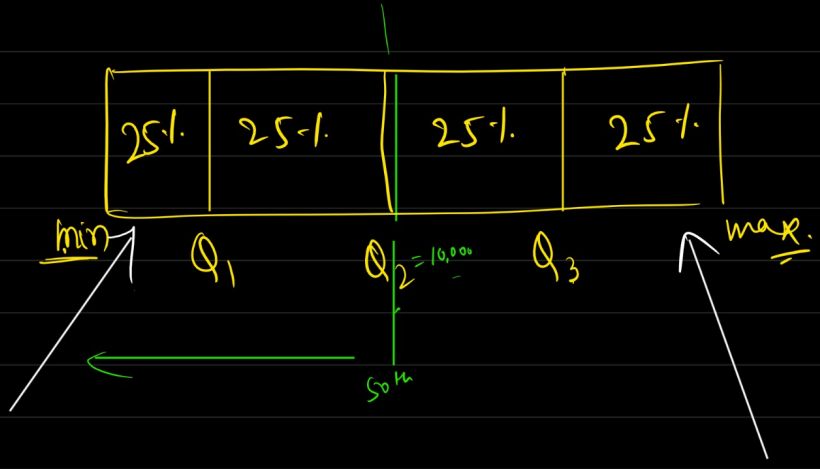
Q_4 (maximum) — 100% percentile

transaction amount
1000
2000
3000
⋮

min (Q_0) → 1000

Q_1 (25^{th} percent) — 5000 ⇒ 25% of transaction amount is equals to or bellow 5000 in the data.

Q_2 (50^{th} percent) — 10,000 ⇒ 50% of transaction amount is equals to or below 10000.



Inter quartile range

$$Q_3 - Q_1$$

2, 3, 3, (3), 3, 4, 4, 5, 5, 5, 6, (6), 6, 7, 8, (99) → outlier

$Q_3 - Q_1$

$25^{th}\% = \frac{25}{100} \times 16 = \frac{1}{4} \times 16^{th} = 4^{th}$ no.

$Q_1 = 3$

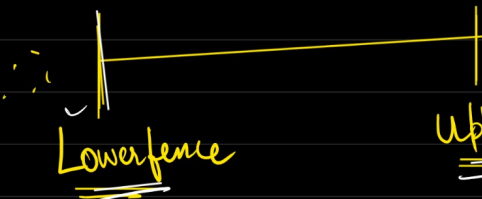
$Q_3 = 75^{th}$ percentile $= \frac{75}{100} \times 16 = \frac{3}{4} \times 16 = 12^{th}$ no.

$Q_3 = 6$

$$IQR = 6 - 3 = 3$$

$Q_3 \quad Q_1$

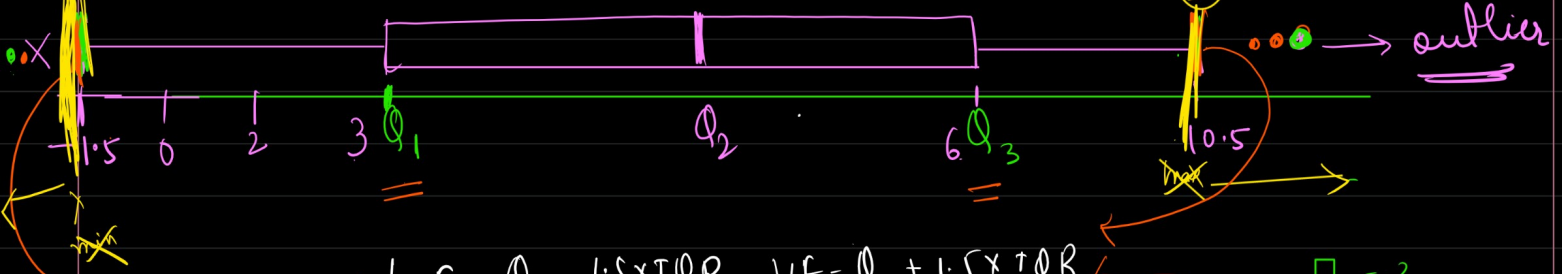
Outliers are extreme values.



Lowerfence $= Q_1 - 1.5\ IQR$

Upperfence $= Q_3 + 1.5\ IQR$

2, 3, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 99



$L.F = Q_1 - 1.5 \times IQR$, $UF = Q_3 + 1.5 \times IQR$

$= 3 - 1.5 \times 3$

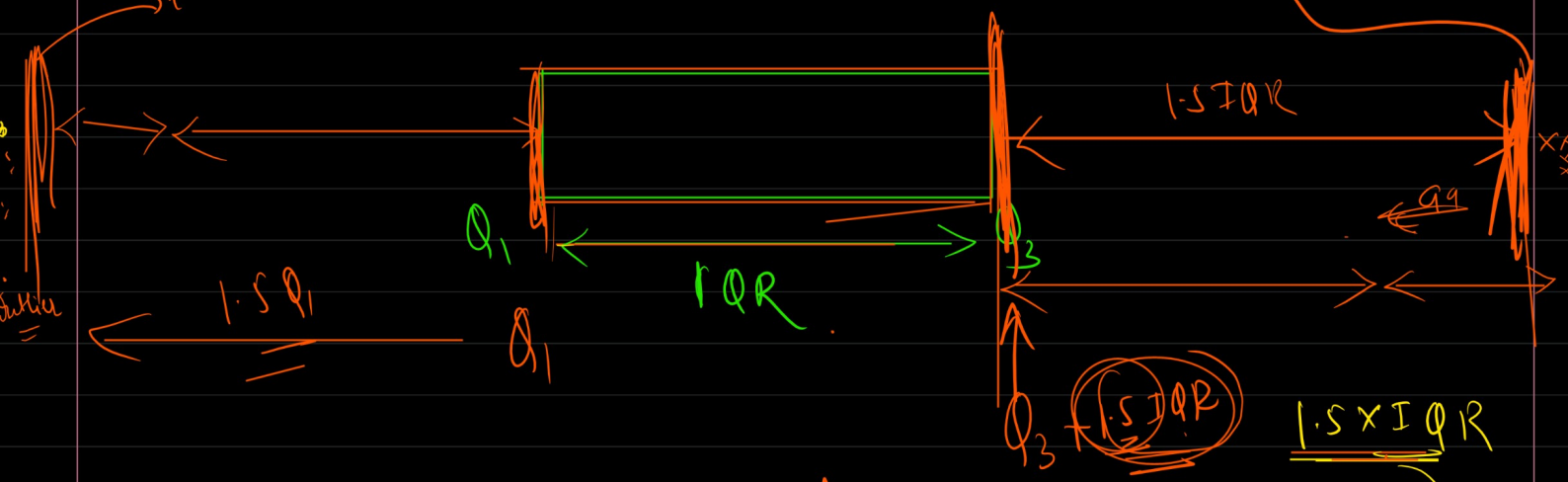
$= -1.5$

$6 + 1.5 \times 3 = 4.5 + 6$

$= 10.5$

$Q_1 = 3$

$Q_3 = 6$



Quartile → Quarter → 4 equal parts.

$1.5 \times IQR$

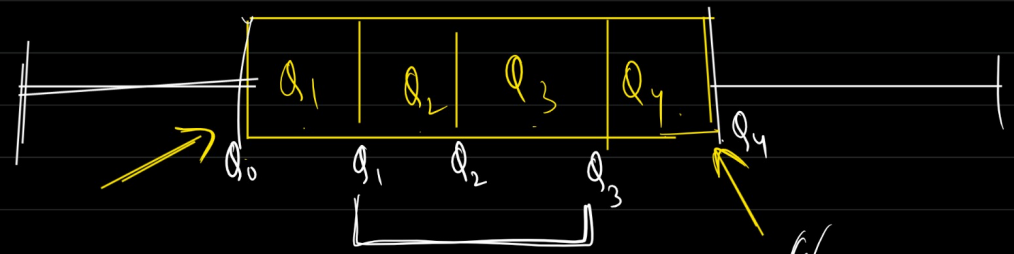
10×1.5

$\Rightarrow 15$

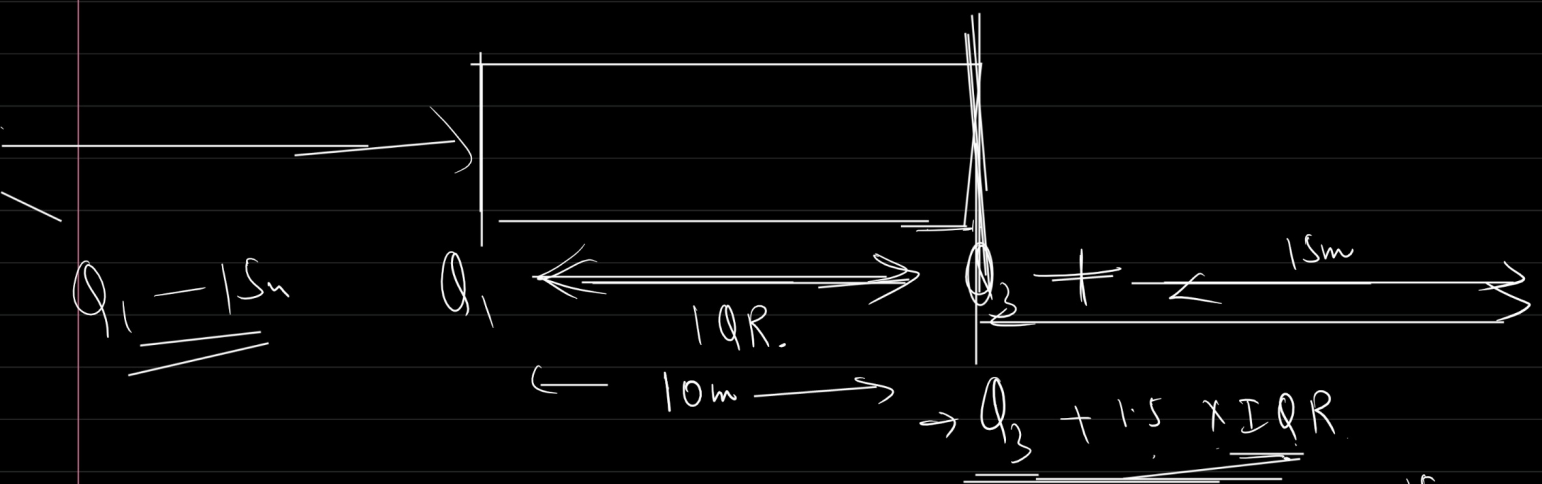
* Measures of Spread | Dispersion

Use — To know outliers

Sns.boxplot( )



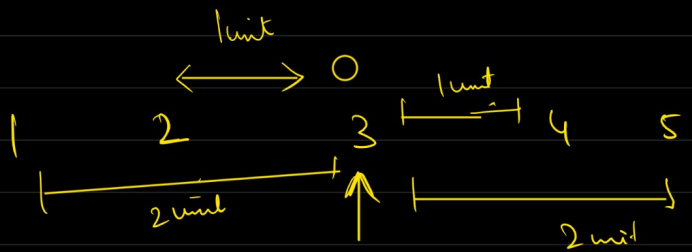
$Q_3 - Q_1 = IQR$ ((Q_0 & Q_4 → is max value))



Measures of spread/dispersion

* Variance
* Standard deviation

* Mean deviation



Mean = 3.

$$\frac{\sum |x - \mu|}{N}$$

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

On an avg each of the data is 1.2 units away from mean value

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

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

↑ Population variance

Sample variance

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

* Variance :-

→ Calculate mean
→ For each no in data, substract the mean and no
→ Square the differnce
→ Calculate the avg of square of difference

→ data 1 2 3 3 4 4 (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

mean = 2.83

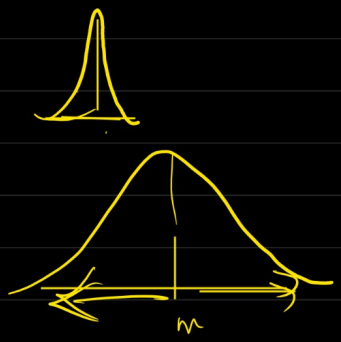
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$= \frac{6.82}{6-1} = \frac{6.82}{5} = 1.37$$

	Population	Sample
mean :	μ	$\bar{x}$
Varian	σ^2	s^2
No of dp	N	n

Spread of data : Variance

Variance ↑ Spread ↑
Variance ↓ spread ↓



* Standard deviation = A measure of how spread the data is

Square root of Variance.

$= \sqrt{\text{Var}}$

Std dev of Pop $= \sigma = \sqrt{\text{Var}_p}$

" " " Sample $= s = \sqrt{\text{Var}_s}$

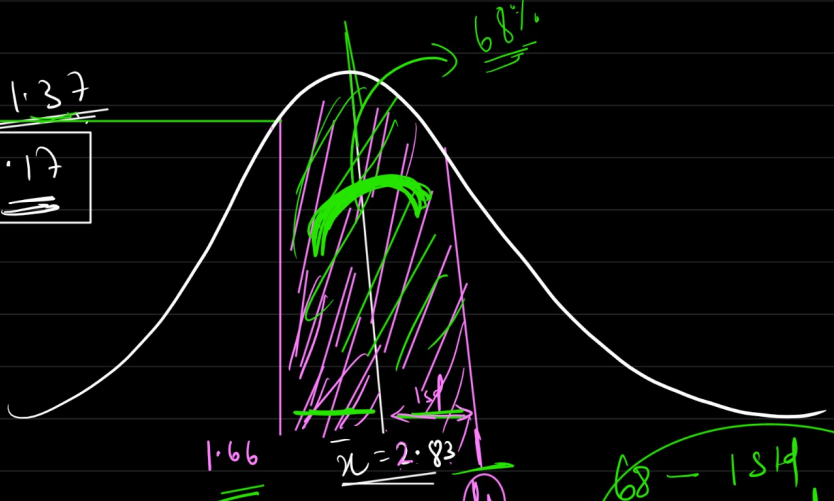
$\sqrt{\text{Var}} = \text{Std}$

- Variance → How much spread.
- Std dev → Standard way of knowing where your data lies

1, 2, 3, 3, 4, 4

→ Var = 1.37

→ std = 1.17



How much dp is 1 sd away from mean

1.17 + 2.83

2.83 + 1.17 / 4

68 — 1 std
95 — 2 std
99.7 — 3 std

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

+3, −3 → $(x-\mu)^2$ → $3^2 + 3^2$

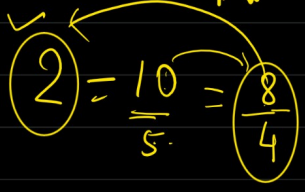
$|x-\mu|$

* Var Sample = $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$ → Bessel correction.

* We are estimating population variance using Sample.

↳ $(x-\mu)^2 > (x-\bar{x})^2$

$(x-\mu)^2 \approx \frac{(x-\bar{x})^2}{n-1}$



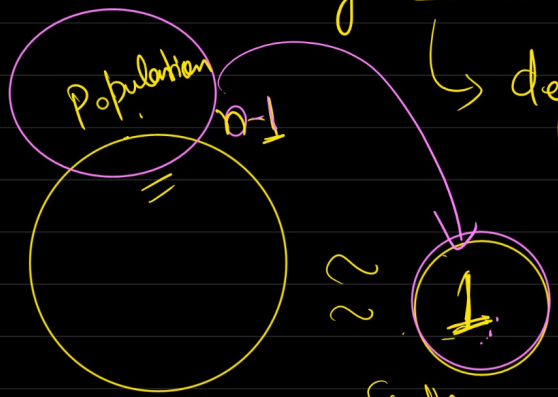
$n-1$ rather than n because Sample variance will be unbiased estimate

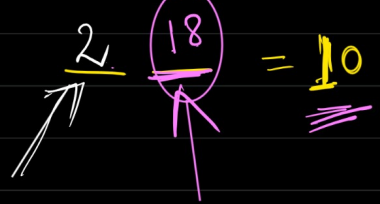
$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Why $n-1$??

↳ degree of freedom.

$df = n-1$



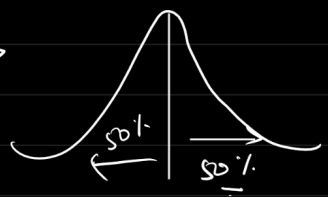


* Measures of Symmetricity

* Skewness

① MCT ✓
② MD ✓
③ MS

Symmetric →



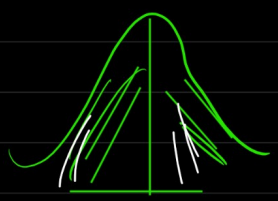
Skewness ⟹ measure of data Symmetry.

$$\text{Skewness} = \frac{x_i - \mu}{n\sigma^3}$$

No Skewness

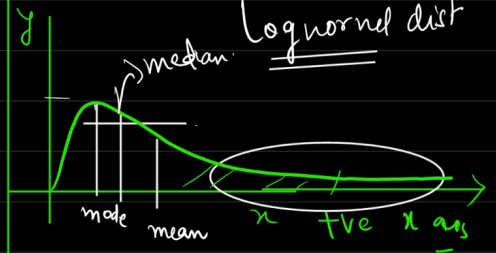
Skewness = 0

Symmetric



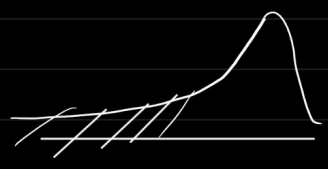
mean = median = mode

Positive skewed



mean ≥ median ≥ mode

Negative Skewed.



mode ≥ median ≥ mean

Q_1 Q_2 Q_3

$Q_3 - Q_2 \simeq Q_2 - Q_1$

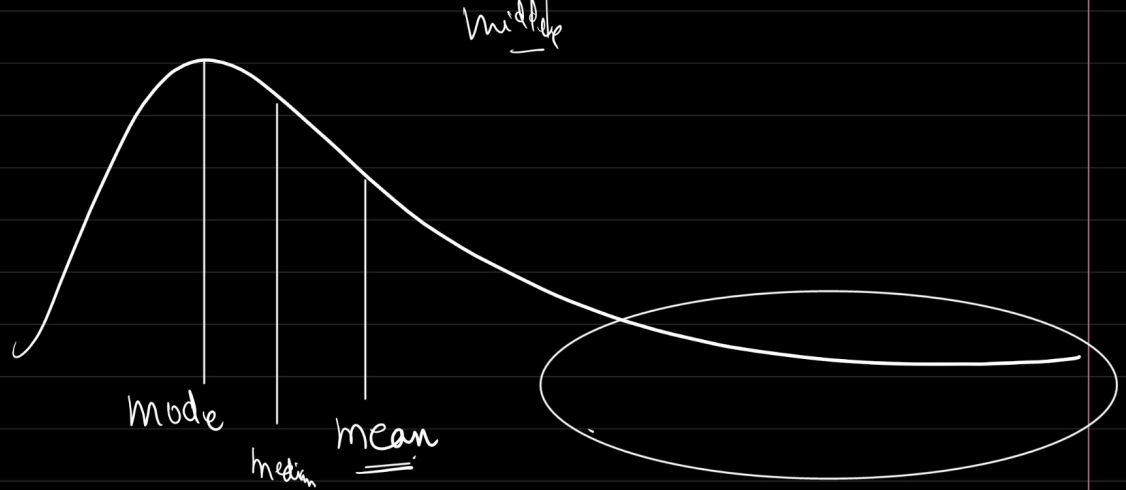
$\longrightarrow$ $Q_3 - Q_2 \geq Q_2 - Q_1$

Transformation

→ log transf
→ reciprocal
– Square
– Square root

Set

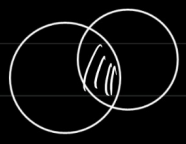
11 — Box Cox transform



x	reciprocal	log	Square	Square root	Box Cox
1	1/1	log(1)	1^2	$\sqrt{1}$	
2	1/2	log(2)	2^2	$\sqrt{2}$	
3	1/3	log(3)	3^2	$\sqrt{3}$	
4	1/4	log(4)	4^2	$\sqrt{4}$	
5	1/5		52		

Set.

① Intersection $A \cap B$

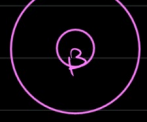


② Union $A \cup B$

③ Difference → items which is present only in first set

$A - B$

④ Subset — All the elements of B present in A



⑤ Subset

⑥ Symmetric diffem →

* Covariance & Correlation