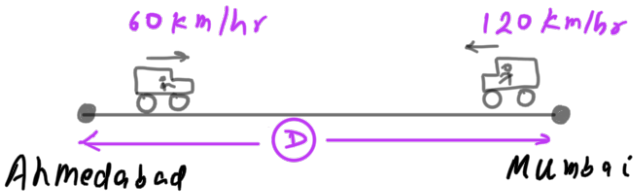
Session 1.3-Statistical Thinking

Table of Content

* Harmonic Mean :-

Eg :-



$$\text{Speed} = \frac{\text{Distance}}{\text{Time}}$$

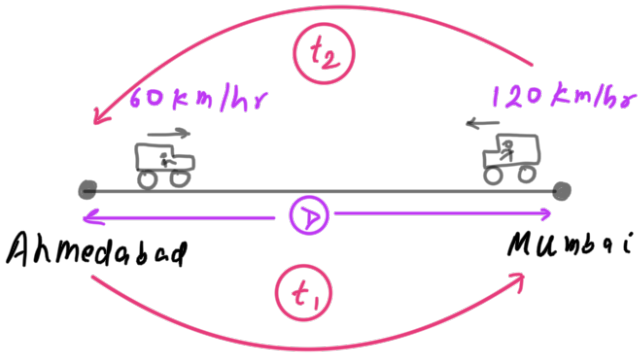
Average speed of car ??

Method 1

A.M. of 60 km/hr & 80 km/hr

$$v_1 = \frac{60 + 120}{2} = \frac{180}{2} \text{ km/hr} = 90 \text{ km/hr}$$

Method 2



i) Ahm. to Mum.

$$t_1 = \frac{D}{60}$$

ii) Mum to Ahm

$$t_2 = \frac{D}{120}$$

$\therefore$ Overall speed $= \dfrac{\text{Total Dis.}}{\text{Total Time taken}} = \dfrac{D+D}{t_1+t_2}$

$$= \frac{2D}{\frac{D}{60}+\frac{D}{120}} = \frac{2}{\frac{1}{60}+\frac{1}{120}}$$

$$= \frac{2 \cdot 60 \cdot 120}{60+120} = 80 \text{ km/hr}$$

~~Method (1) → 90 km/hr~~ ✗

Method (2) → 80 km/hr

So, H.M.

$x_1, x_2, x_3, \ldots, x_n$

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} 1/x_i}$$

https://stats.stackexchange.com/questions/23117/which-mean-to-use-and-when

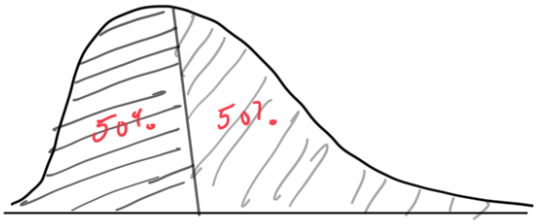
* Quartiles, Quantiles, Percentile

Quartile

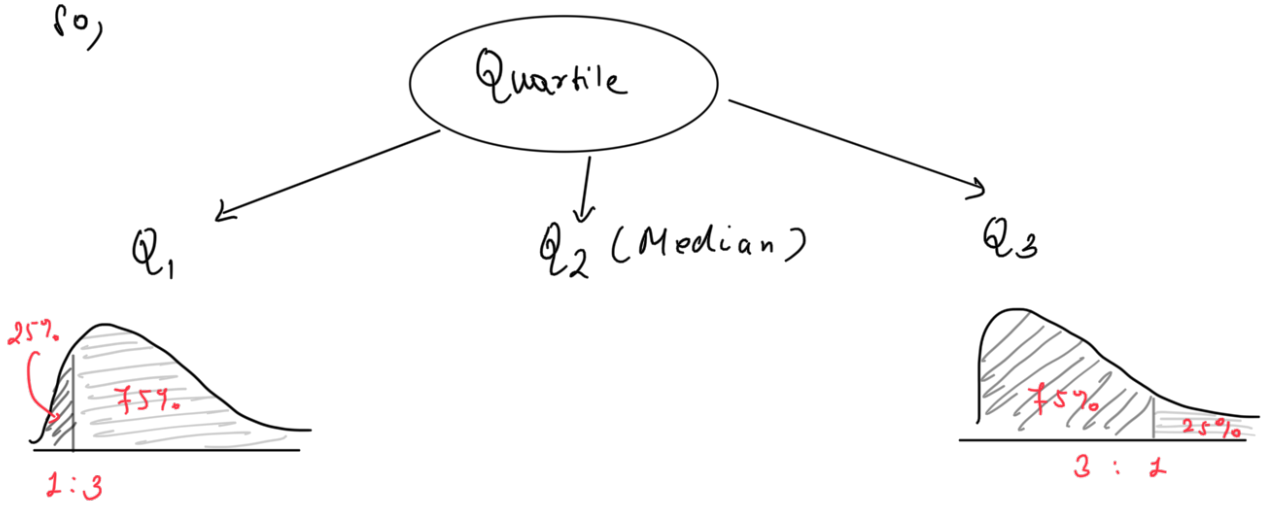
↓

Divides the frequency Distribution into equal parts.

Recall Median



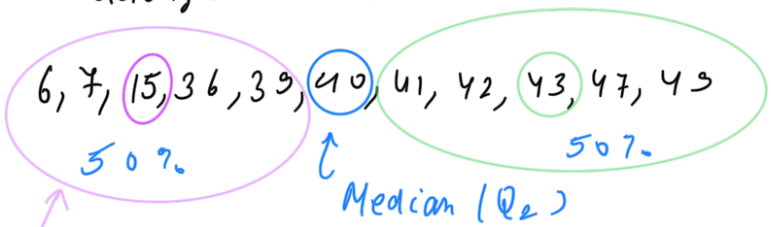
So,



Eg:: 6, 15, 39, 41, 43, 45, 7, 36, 40, 42, 47

Q_1, Q_2, Q_3 = ??

Soln: Arrange in ↑ order

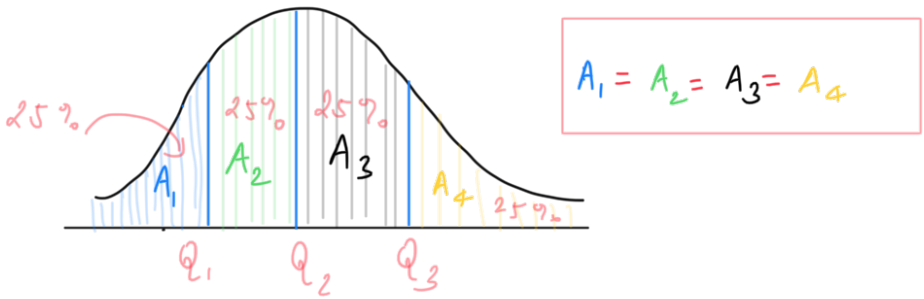


Q1: It is just a median of [minimum, Q_2]

Q3: It is just a median of [Q_2, maximum]

Note:- Follow Wikipedia

* Percentile :-



for percentile divide the entire data into 100 equal parts.

Eg:- $P_{40} = \alpha$

↳ 40 % of the data is below α.
{ Doesn't mean less than α }

Counter Example :- 10, 10, 10, 10, 10, 100

Note:- Percentile Rank is Different thing.

* Quantiles :-

Divide into arbitrary proportion.

Eg:- $q(0.312)$

* Measures of Dispersion

- Why is it required ?

Ex:

Income in (k)

FAMILY (A) : 20, 30, 40, 50, 60, 70
FAMILY (B) : 20, 43, 44, 46, 47, 70
FAMILY (C) : 40, 43, 44, 46, 47, 50

} some has extreme wealth in family.

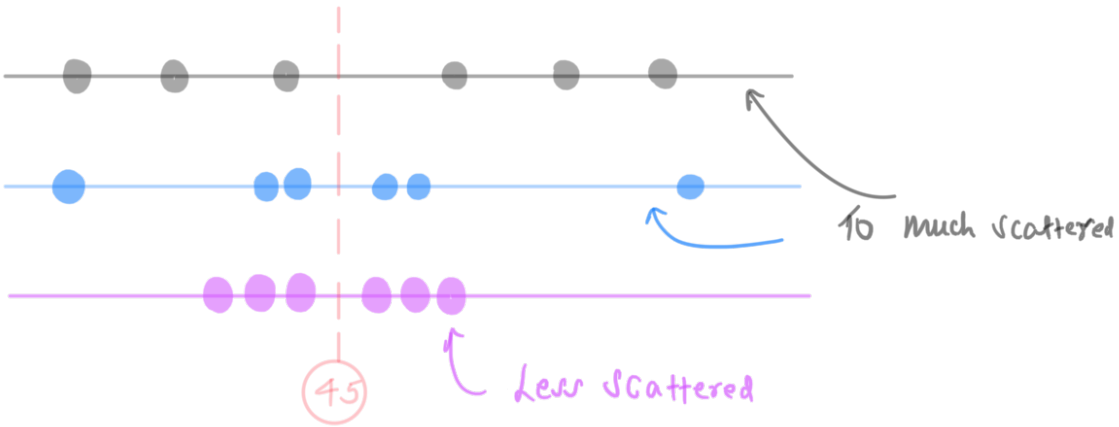
$\bar{x}_A = m_A = 45$
$\bar{x}_B = m_B = 45$
$\bar{x}_C = m_B = 45$

All have same centre but data is different !

Measure of centre Alone can be misleading

Comparision b/w Data sets is lost !

Pictorial



* Dispersion tells us how the obr^{n}s in a data set vary among themselves.

* RANGE :-

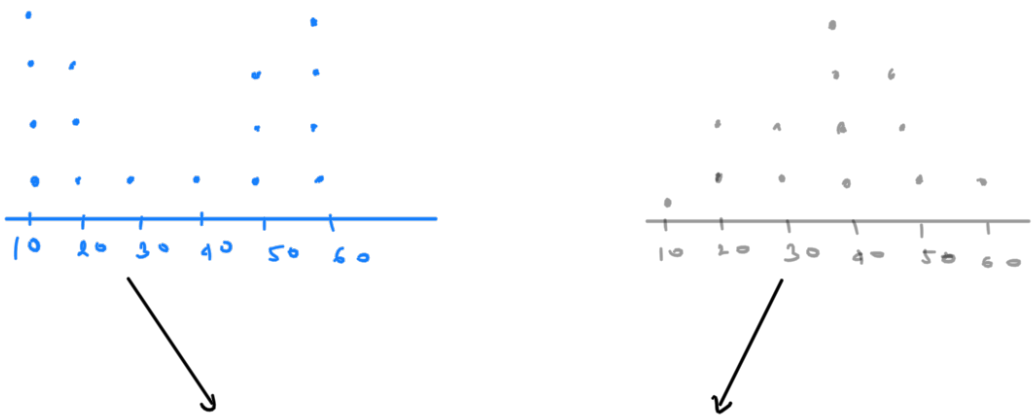
- Simplest measure of Dispersion

Range = Max - Min

Eg:- 2, 4, -8, 10, 4, 14

Range = 14 - (-8) = 22

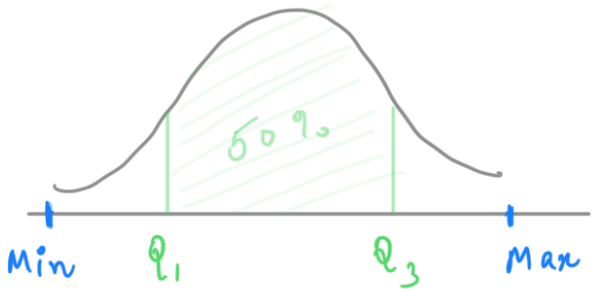
* Drawbacks of Range



Same Range, But Different pattern of Variability

* Range does not consider the form of dist^{n}.
* Range is highly affected by presence of single outlier
* Range is highly affected by sampling fluctuation.

* IQR { Inter Quartile Range }



$$IQR = Q_3 - Q_1$$

Not influenced By extreme values

IQR is useful to detect outlier in Data set

Standard Rule for Outlier

If $x_i \geq Q_3 + 1.5(IQR)$

or

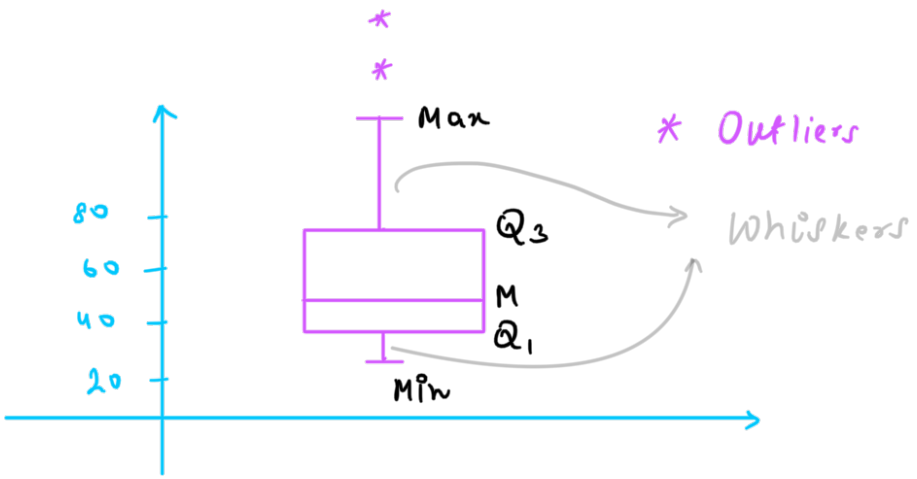
$x_i \leq Q_1 - 1.5(IQR)$

* Five Number Summary :-

Min	Q_1	Med (Q_2)	Q_3	Max

i) Median describes the centre of distⁿ
ii) Quartiles show the spread of the central half
iii) Min & Max show the full spread of Data.

↘ Visual Repr^{n}



BOX PLOT says a lot about Data

* Detailed discussion will be later on Box plot.

* Go to Notebook

* Mean Deviation :

  * Why is it required ?
  → Range, IQR these all measures do not take into account all obsn^{s} in the Dataset.

* A good measure of variability should depend on each obsn.

$x_i - \bar{x}$

Deviations : How much an observations is far from its central value

How to combine all deviations into single numerical measure?

* Take average of deviations?

$$\frac{1}{n}\sum(x_i-\bar{x}) = ?$$

* Take absolute deviations

$$MD_{\bar{x}} = \frac{1}{n}\sum|x_i-\bar{x}|$$

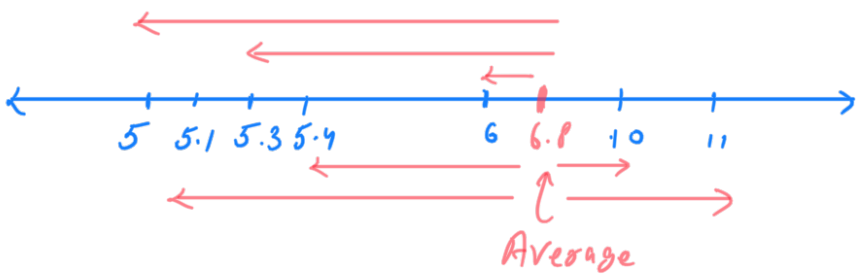
On an average, how much a value is far from its central value.

* Variance:

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

Visualization

Eg: 5, 5.1, 5.3, 5.5, 6, 10, 11



** Squaring terms impact more than linear

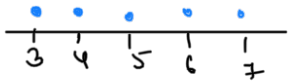
* Standard Deviation

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

Taking square root give same unit.

Ex: D_1: 3, 4, 5, 6, 7 $\qquad$ D_2: −8, 0, 5, 10, 13

mean = 5 $\qquad$ mean = 5



$\sigma = 1.414$ $\qquad$ $\sigma = 7.45$

D_2 has high variability (more) w.r.t. D_1

* Popⁿ (parameter) $\qquad$ Sample (statistic)

mean : μ $\qquad$ mean : $\bar{x}$

variance : σ^2 $\qquad$ variance : $s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1}$

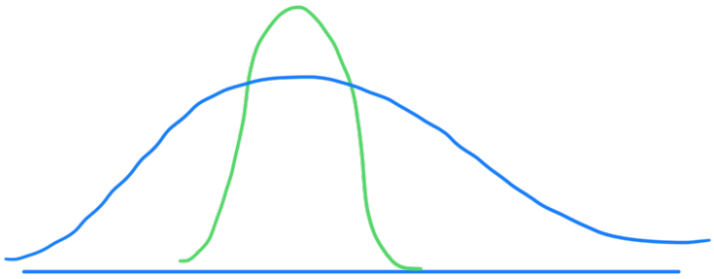
std : σ $\qquad$ std : s

$\bar{x}$ is an unbiased estimator of μ

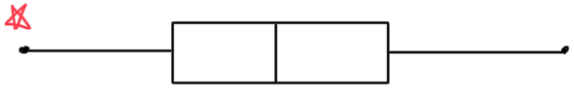
s^2 is an unbiased estimator of σ^2

while s is not an unbiased estimator of σ

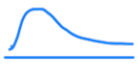
Que



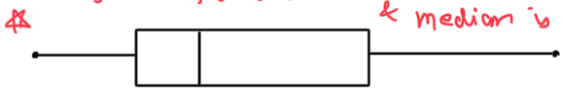
* Box Plot.



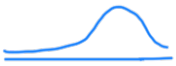
Normal Dist^n



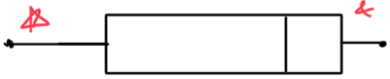
for left side → whisker length is short & median is on left



Positive skew



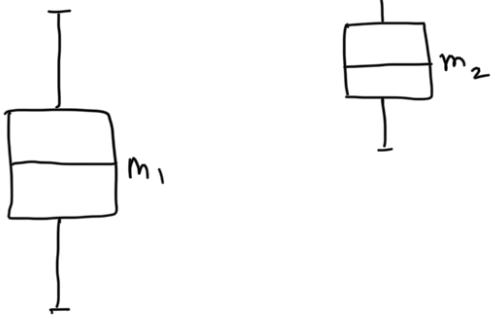
for right side → whisker length is short & median is on right



Negative skew

* Comparision of Box Plots :

i> Compare the medians of the box plot



$m_1 < m_2$ ⇒ there is likely to difference b|w groups.