Day-2 – Statistics 04/09/2022

1) Histogram:

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 68, 78, 90, 95, 100}

① Sort the Numbers
② Bins → No. of groups
③ Bins size → size of Bins

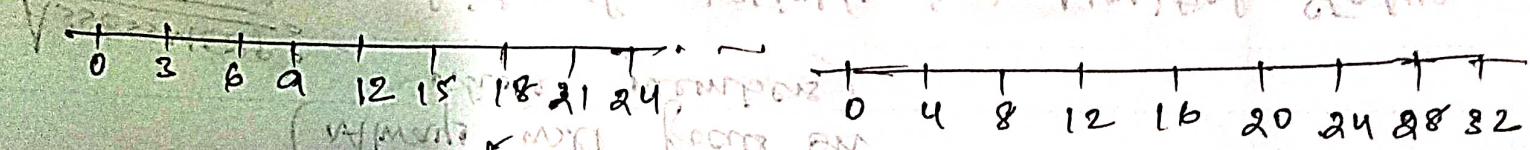
min = 10
max = 40

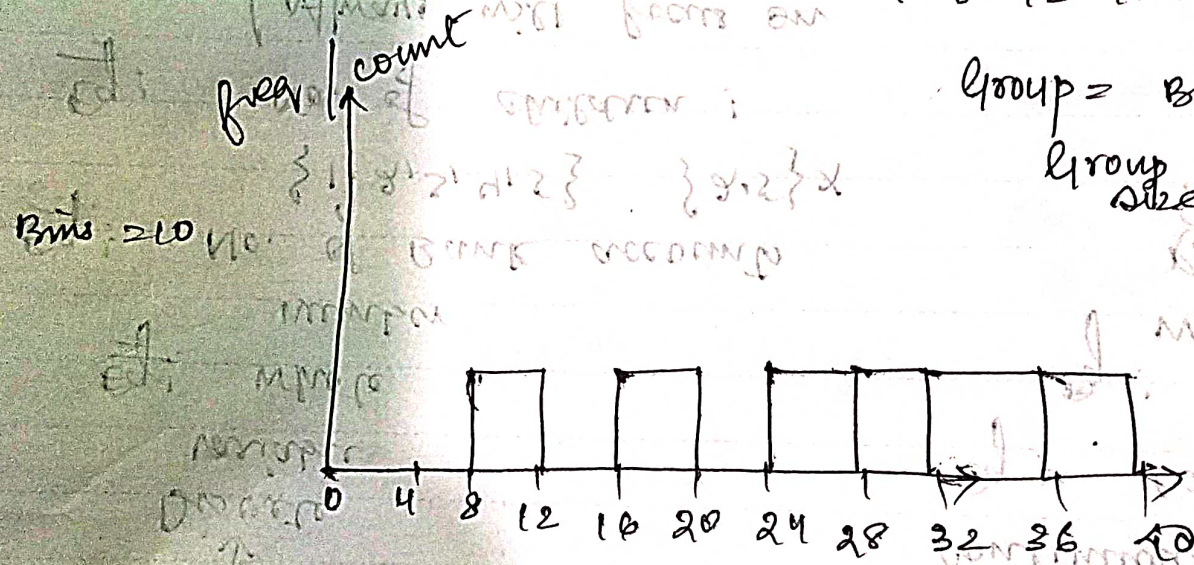
[10, 20, 25, 30, 35, 40] → Not considering

Bin Size = Max / Bins

Bins = 10 → 40/10 = 4

(40−10)/10 = 3





Group = Bins

Group size = Bin size

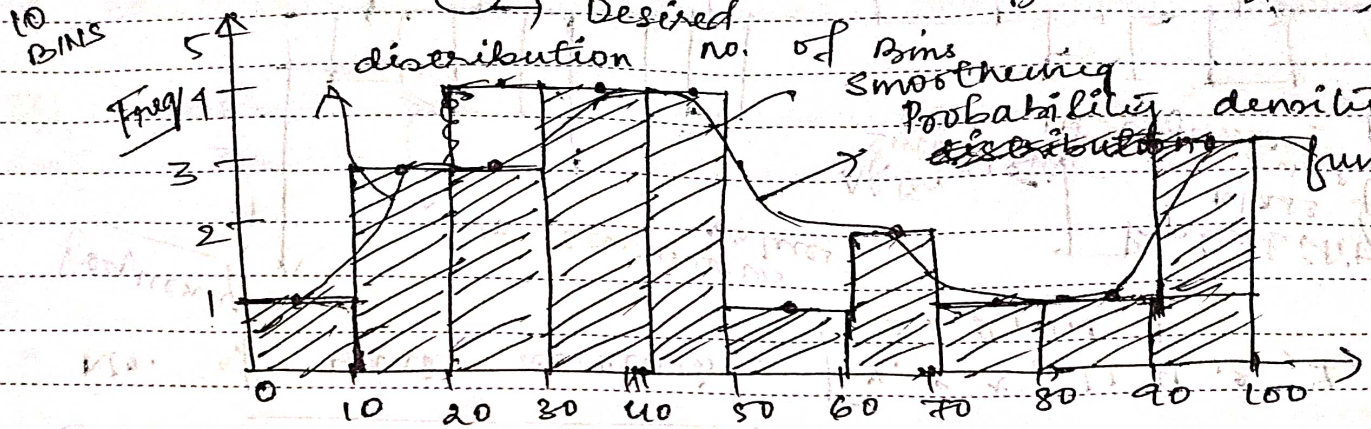
Bin size → $\frac{100}{10} = 10$.

10 BINS

Desired distribution

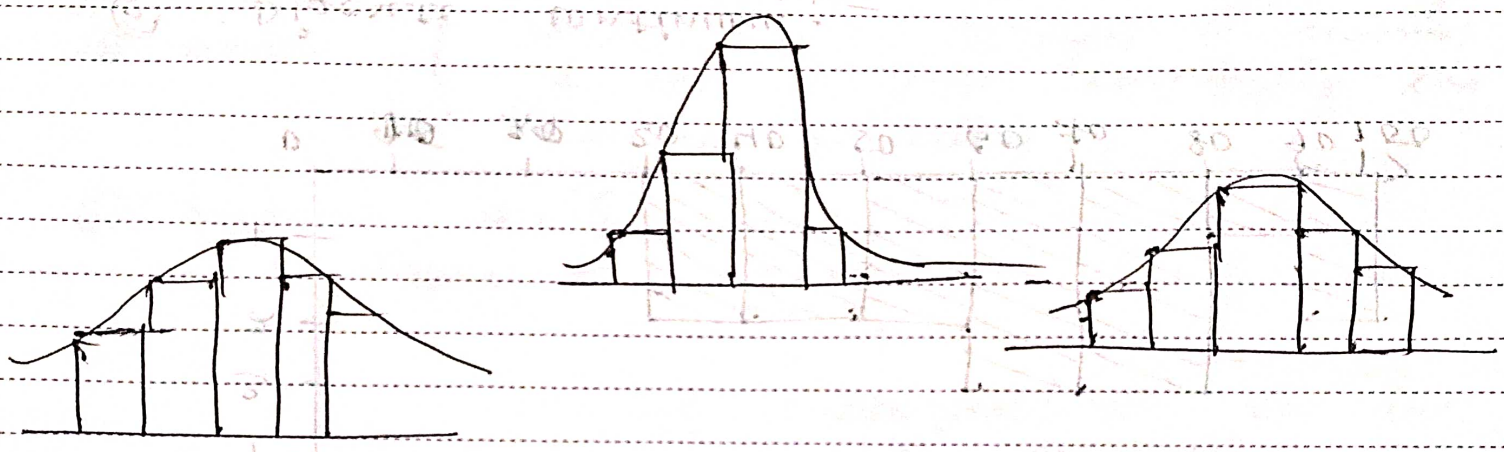
20 BINS → $\frac{100}{20} = 5$

no. of Bins



Smoothening

Probability density func.



Smoothening — Helps us to smoothen the histogram by drawing the line, ~~and~~ creating a distribution, and hence the probability density function (PDF).

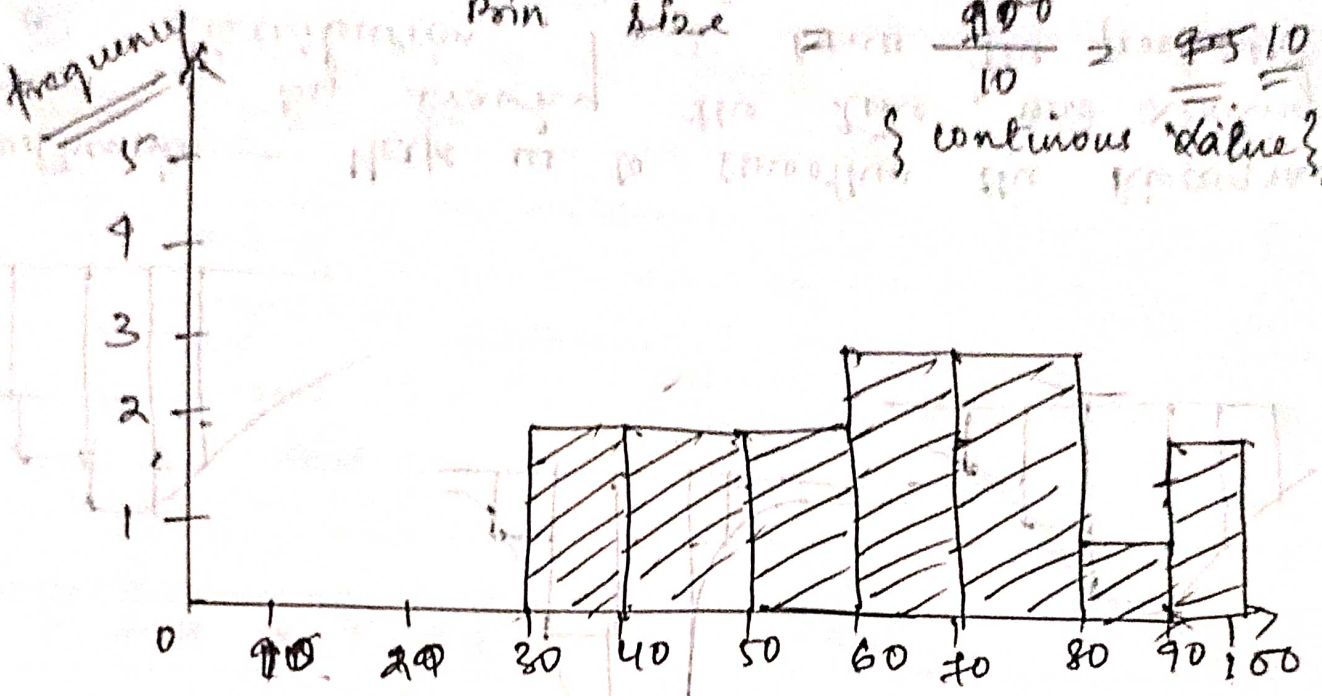
Weight = { 30, 21, 25, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95 } 100 }

Bins = 10 { Bin size = $\frac{95-30}{10} = \frac{65}{10} = 6.5$
→ Not starting from 0 }

If starting from 0, then,

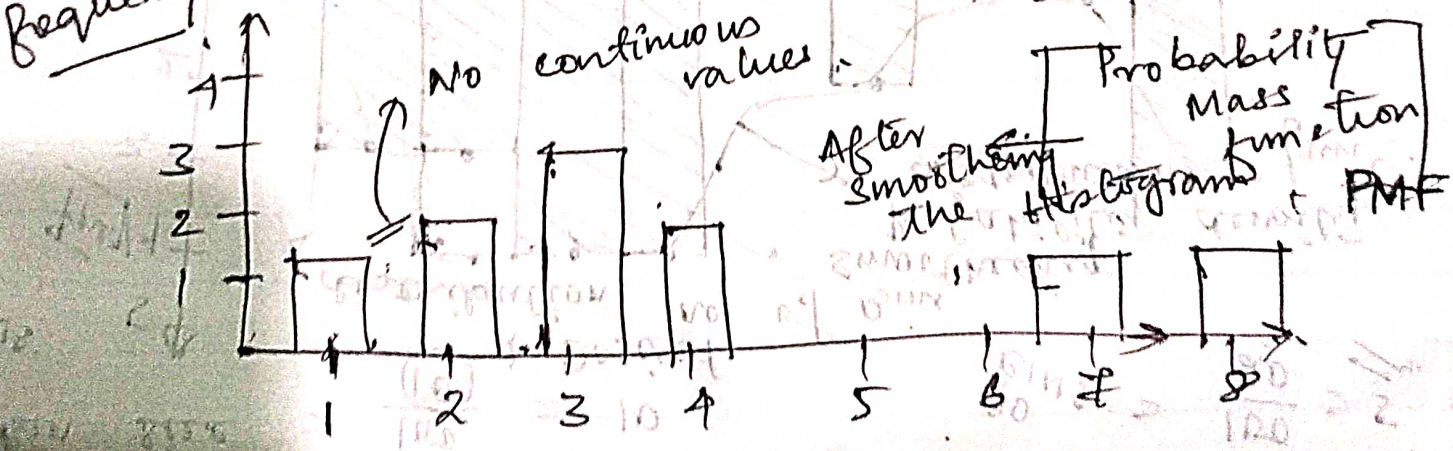
Bin size = $\frac{100}{10} = \frac{95}{=} \frac{10}{=}$

{ continuous value }



② Discrete continuous :-

No. of Bank accounts = [2, 3, 5, 1, 4, 2, 7, 8, 3, 2, 4, 5]



→ Discrete has whole number values.
→ continuous have decimal and whole numbers.

Pdf: Probability density function → Continuous
Pmf: Probability Mass function → discrete

* II Measures of Central Tendency:

① Mean
② Median
② Mode

(A measure of cTendency is a single value that attempts to describe a set of data identifying the central position)

① Mean = $x = \{1, 2, 3, 4, 5\}$

Avg (Mean = $\frac{1+2+3+4+5}{5} = 15/5 = 3$.

Centroid

element in the set of elements.

Population (N).

Population mean (μ)

$$\mu_2 \quad \sum_{i=1}^{N} \frac{x_i}{N}$$

Sample (N) ~~mean~~

Sample mean ($\bar{x}$)

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

Ex: Population = {24, 28, 21, 28, 27}
Age

Population mean (μ) = $\frac{(24+23+21+28,27)}{6}$ = 17.5

Sample mean: ($\bar{x}$) = $\sum_{i=1}^{n} \frac{x_i}{n}$ = $\frac{24+27+1+2}{4}$

$\bar{x}$ = 13.5

$N > n$ → for more all cases → this cannot happen

$\mu > \bar{x}$

$\bar{x} > \mu$ → this can happen

✓ Practical Application : (Feature Engineering)

Age	Salary	Family size
NAN		NAN
	NAN	
NAN	NAN	NAN

dropping the row / columns may lead to info loss.

So, we can take the mean of the columns and substitute the value with

the missing value. Because, mean represents the value of central tendency.

② ~~Median~~: EX:

Age	Salary
24	45
28	50
29	NAN → Salary 62
NAN → 29.6 → 38	60
31	75
36	80
NAN	NAN
80 ← outliers	200

Because of the outliers, the mean value changed.

No of matching values.

To solve this problem, we apply median.

(If mean is changing drastically ~~with~~ then there must be an outlier present in the column).

② Median:

Steps to find out median:

1) Sort the Numbers
2) Find the central Number.

1) { if the no. of elements are even, we find the avg. of the central elements }.

2) if the no. of elements are odd, we find the central elements.

Eg:

{ 0, 1, 2, 3, 4, [5, 6], 7, 8, [100, 120] } → Avg.

Median = $\frac{5+6}{2} = 5.5$,

Mean = 25.6

Median = 5

No outliers ⇒ Mean

with outliers → Median.

No such diff.

② Mode: { Most frequent occurring elements }

Eg: { 1, 2, 2, 3, 3, 4, 3 } = Mode = 3

Application:

Types of Flowers

daily
Sunflower
Rose
NAN ← Rose
Rose
Sunflower
NAN ← Rose

{ Replace the NAN value with max. no. of occurence here, Rose }

III Measures of Dispersion :-

1) Variance (σ^2) ← Spread of data
2) Standard Deviation (σ)

Variance

population variance (σ^2)

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

$x_i - \mu$ = Distance from the mean.

Sample variance (s^2)

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

Why $n-1$? (Assesment)

Variance

$\{1, 2, 3, 4, 5\}$

$\mu = 3$

$$\sigma_2^2 = \frac{\left[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2\right]}{5}$$

$$= \frac{4+1+0+1+4}{5} = 10/5$$

$$= 2$$

variable

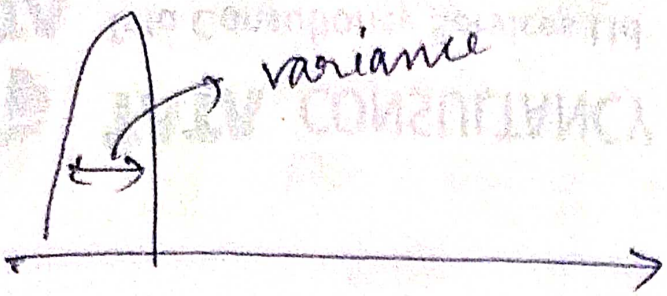
$\{1, 2, 3, 4, 5, 6, 80\}$

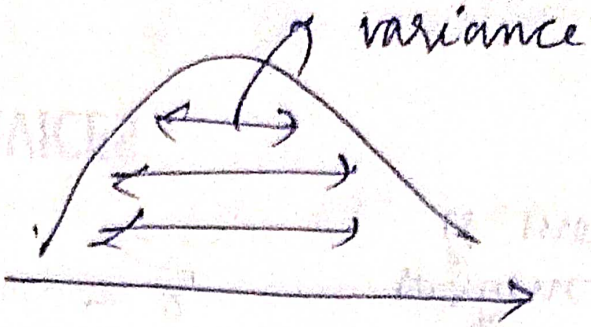
$\mu = 14.4$

$\sigma^2 =$ ____

$\sigma^2 = 419.4$

variance is more.

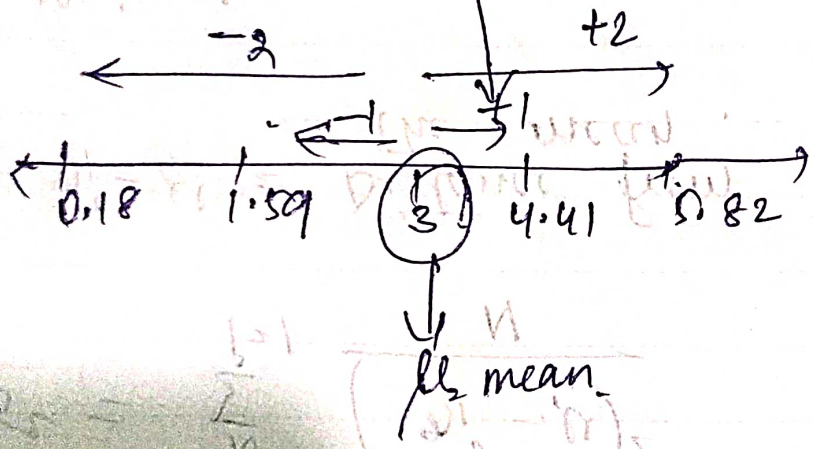




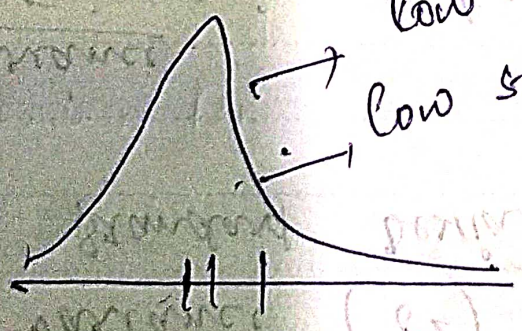
(4) Standard Deviation: $(\sqrt{\sigma^2})$

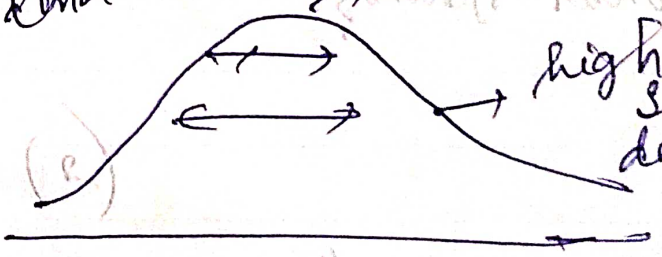
$\{1, 2, 3, 4, 5\}$ = $\frac{15}{5}$ 23

$\sigma^2 = 2$ $\sigma = 1.414 = \sqrt{2}$



How many std. deviation 4 away from the mean? ⇒ +1





III Percentiles and Quartiles:

Percentage = {1, 2, 3, 4, 5, 6, 7, 8}

Percentage of even no. = $\frac{\text{No. of even nos.}}{\text{Total no. of Num:}} = \frac{4}{8} = 0.5 = 50\%$.

Percentile = Rate, CAT, JEES, etc.

Defn:- A percentile is a value below which a certain percentage of observation lie.

Ex: 99 percentile: It means the person has got better marks than 99 other students

Dataset: {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12} → should be ascending.

What is the ~~percentage~~ percentile of rank 10?

Percentile rank of x → $\frac{\text{No. of value below } x}{n}$

$= \frac{16}{20} = 80$ percentile

Q: what is the value that exists at 25 percentile?

$$\text{value} = \frac{\text{Percentile}}{100} \times n$$

$$= \frac{25}{100} \times 20$$

$$= 5^{th} \text{ index}$$

value is at the 5th index.

Index starts from 0.

$\frac{95}{100} \times 20$

$= 19^{th}$ index

for even $= x(n+1)$

for odd $= x(n)$

5 Number Summary:

1) Minimum
2) first Quartile
3) Median
4) Third Quartile
5) Maximum.

Use to remove the outliers.

Eg: {1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 9, 27}

0 → 5.25 index

Thus, we try to create a fence.

15.75 index

[Lower fence ⟷ Higher fence]

Lower Fence $= Q_1 - 1.5 \cdot (IQR)$

[−3.65 ⟷ 14.25]

Lower fence

Higher fence

27, is not in the fence, so this is outlier.

Lower Fence = $Q_1 - 1.5(IQR)$

Higher Fence = $Q_3 + (1.5)(IQR)$

$IQR = Q_3 - Q_1$

$75 \quad 25$

$\Rightarrow 7.5 - 2 = 4.5$

Inter Quartile Range.

(Basically, the std. deviation after several exp.)

$Q_1 = \frac{25}{100} \times (n+1)$

$= \frac{25}{100} \times 21$

$= 5.25 \Rightarrow$ Index.

$5 \rightarrow$ 1: there are no 5.25 Index, so avg $\left(\frac{3+3}{2} = 3\right)$

$Q_3 = \frac{75}{100} \times 21 = 15.75$

15th index $+ \frac{8+7}{2} = 7.5$

Thus,

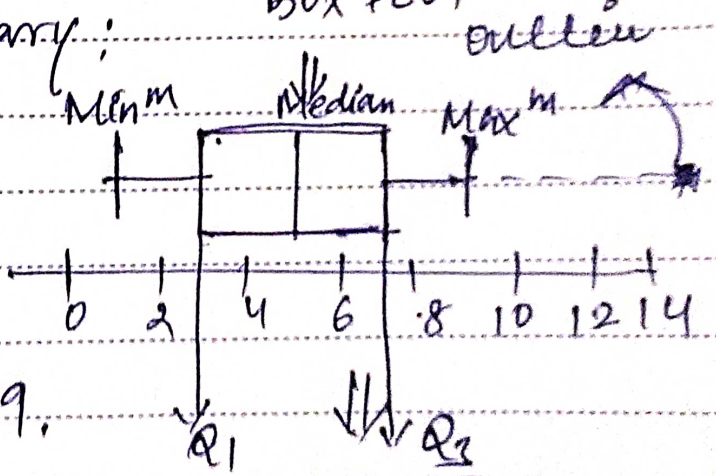
Lower fence $= 3 - 1.5 \times 4.5 = -3.65$

Higher fence $= 7.5 + (1.5)(4.5) = 14.25$

Thus, 5 number summary:

1) Minimum = 1
2) $Q_1 = 3$
3) Median = 5
4) $Q_3 = 7.5$
5) Maximum = 9.

BOX PLOT



To treat outliers.