Basic Stats

1) Histogram

A histogram is a graphical representation of a grouped frequency distribution with continuous classes. It is an area diagram, and can be defined as set of rectangles with bases along with the intervals between class boundaries and its area proportional to frequencies in corresponding classes.

Example: Lets consider Age dataset.

Ages = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100 }

* Steps to create / Draw Histogram

1) Sort the dataset in Ascending order.

2) Bins : No. of groups 3) Bin size : size of Bins.

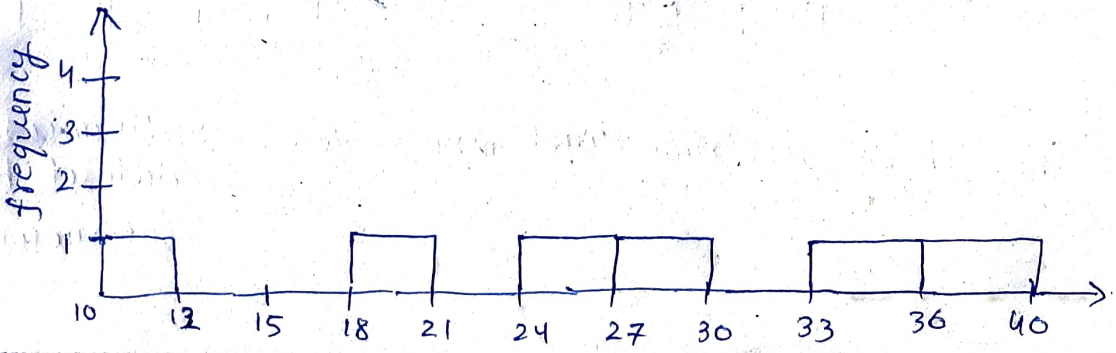
Eg: data = { 10, 20, 25, 30, 35, 40 }.

* we want 10 bins. (say).

min = 10 , max = 40.

$$\text{Bin size} = \frac{\text{max} - \text{min}}{\text{bins}} = \frac{40-10}{10} = \frac{30}{10} = 3$$

So we will have 10 bins with Bin size of 3.

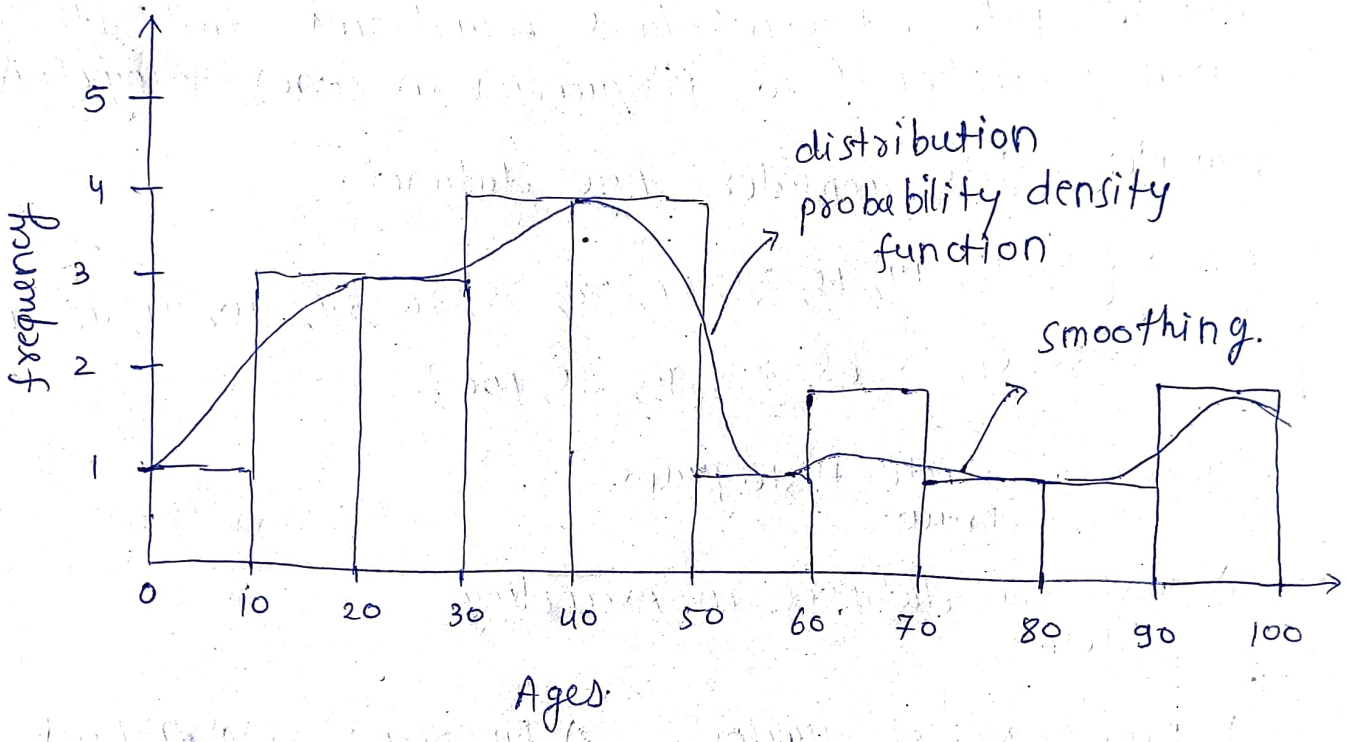


for Age dataset

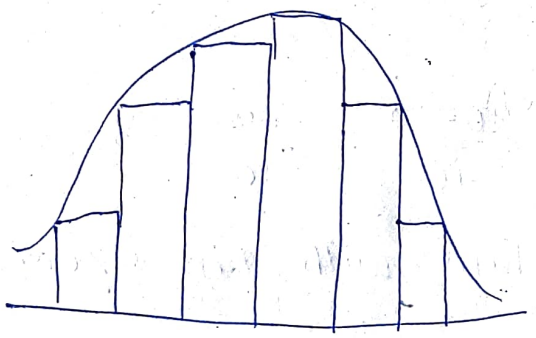
Age = { 0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100 }.

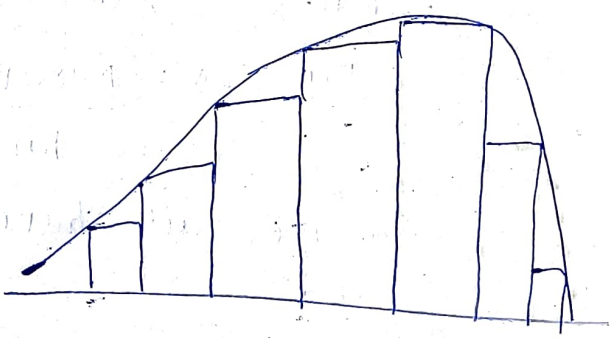
min = 0 max = 100 bins = 10 (say)

∴ Bin size = $\frac{100-0}{10}$ = 10



* smoothing will help to form the distribution as shown below:

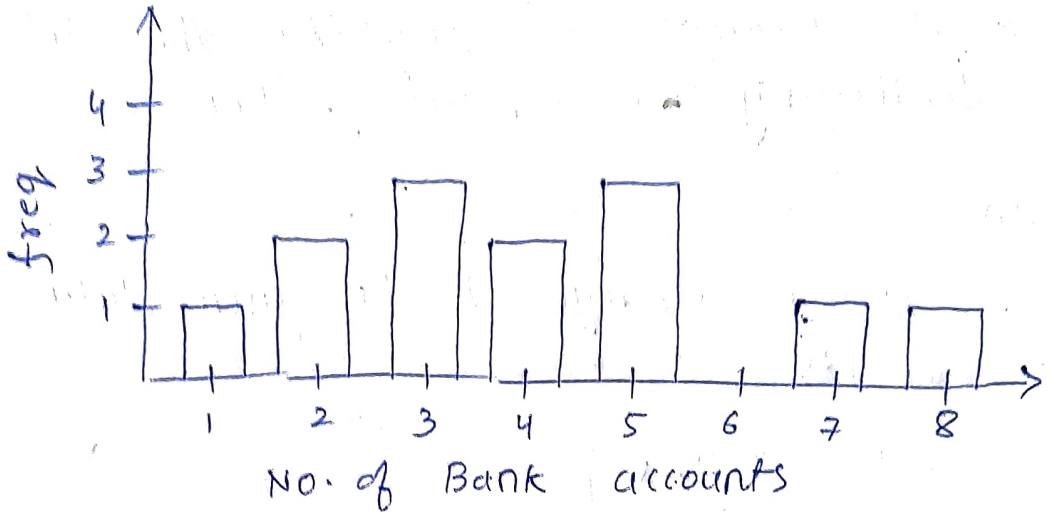




The above distributions are for continuous values. (dataset having continuous values).

Discrete continuous

No. of Bank accounts = {2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5}

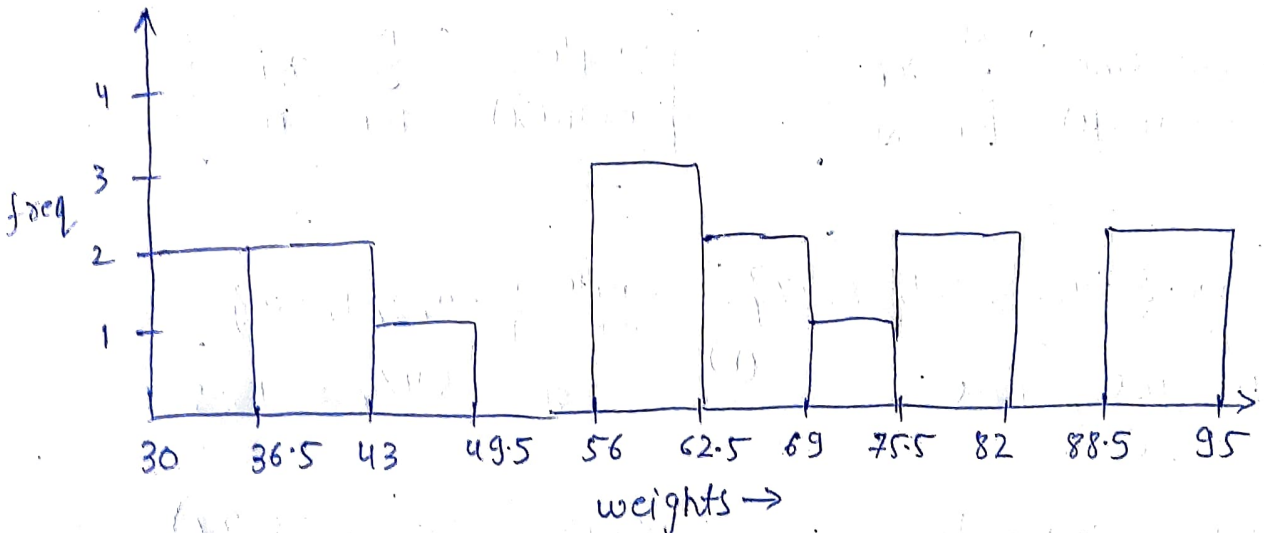


eg: weights = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}.

Bins = 10

min = 30 max = 95

Binsize = $\frac{95-30}{10} = 6.5$



Note 1) for smoothing the continuous histogram we will use probability density function (Pdf)

2) For smoothing the discrete continuous histogram we will use probability mass function (Pmf)

2) Measure of Central Tendency.

a) Mean b) Median c) Mode

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

① Mean

Mean is the average of the given data-set elements.

Ex: Rank = {1, 2, 3, 4, 5}.

$$\text{Mean} = \frac{\text{Sum of all elements}}{\text{No. of elements}} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

population (N) sample (n)

(N >>> n.)

$$\text{population mean } (\mu) = \frac{\sum_{i=1}^{N} x_i}{N}$$

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

Eg:

popln age = {24, 23, 2, 1, 28, 27} ①

$\mu = \frac{105}{6} = 17.5$

sample age = {24, 2, 1, 27}

$\bar{x} = 54/4 = 13.5$

Note: N > n ②

sample age = {24, 23, 28, 27)

$\bar{x} = 102/4 = 25.5$

$\mu > \bar{x}$ ① $\bar{x} > \mu$ ②

* practical application of mean (feature Engg)

Age	salary (K)	No. of Invest-ment
18	33	2
20	42	NaN
NaN	50	3
22	NaN	NaN
33	20	1
29	NaN	4
NaN	27	2
30	40	NaN

← if we drop rows containing NaN (null values) then there will be loss of data.

So better option is to Replace NaN values By mean of Age, salary ....

$\mu_{Age} = \frac{18+20+22+33+29+30}{6}$ (we ignore NaN values for mean calculation)

$= \frac{144}{6} = 24.$

$\mu_{salary} = \frac{33+42+50+20+27+40}{6} = 35.33$

Now let's introduce an outlier in above data set.

Age	salary (K)	No. of Investment
90	240	9.

$(\mu_{age})_{New} = \frac{144+90}{7} = 33.42$

$(\mu_{sal})_{New} = \frac{212+240}{7} = 64.57$

* presence of outliers in data set increases/decreases mean value (so to avoid the effect of outliers we use median).

② Median

median is middle number in sorted, ascending or descending dataset elements. It is the point above and below which half (50%) observed data falls. So it represents midpoint of data.

Eg: {1, 2, ③, 4, 5}

$\bar{x} = 15/5 = 3$

median = ③

{1, 2, [3, 4], 5, 100}

↑ outlier.

$\bar{x} = 115/6 = 19.16$.

$\text{median} = \frac{3+4}{2} = 3.5$

∴ It is better to use median in case of outliers

* steps to find out Median.

1) Sort the list in ascending / descending order.

2) Find central Number.

a) if No. of elements are even

Median = Avg. of central No. = $\frac{C_1 + C_2}{2}$

b) if No. of elements are odd

median = central No. = C.

Eg: {1, 2, 3, 4, [5, 6], 7, 8, 100, 120} - Even No's.

$\text{median} = \frac{5+6}{2} = 5.5$

{0, 1, 2, 3, 4, [5], 6, 7, 8, 100, 120} - odd nos.

median = 5

② Mode

It is the most frequently occuring element in dataset (in sorted dataset)

eg: {1, 2, 2, [3, 3, 3], 4, 5} {1, [2, 2, 2], [3, 3, 3], 4, 5}

mode = 3 mode = 2, 3.

* mode is used to handle NaN values in categorical variable

eg:

Type of flower.
Lily
sunflower.
Rose
NaN
Rose
sunflower
Rose
NaN

mode = Rose

∴ NaN values can be replaced by Rose.

③ Measure of dispersion

① variance (σ^2)

② standard deviation (σ)

① Variance

Variance is stastical measurement of the spread between numbers in dataset. More specifically, variance measures how far each number in the set is from [mean], and thus from every other number in set.

population variance (σ^2)

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

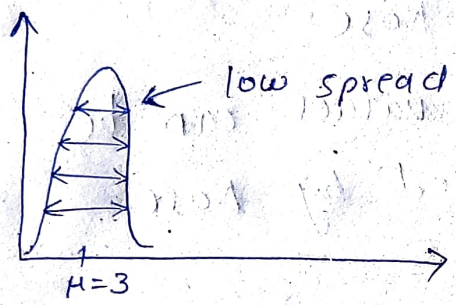
sample variance (s^2)

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

eg: $\{1, 2, 3, 4, 5\}$ $\mu = 3$

$$\sigma^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}$$

$$= \frac{4 + 1 + 0 + 1 + 4}{5} = \frac{10}{5} = 2.$$

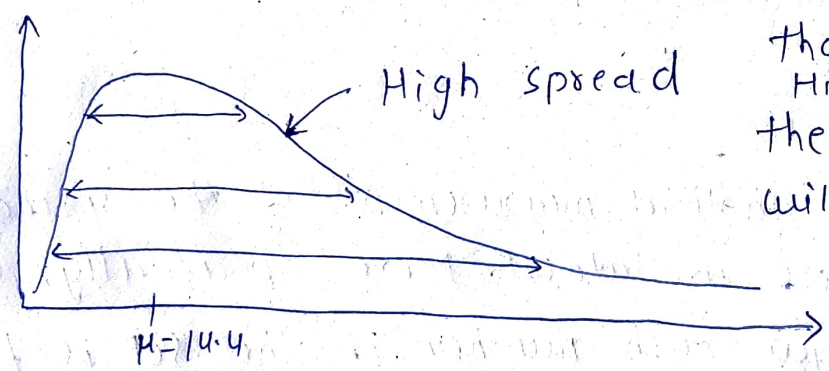


Now consider the same with an outlier

$\{1, 2, 3, 4, 5, 6, 80\}$ $\mu = 14.4$

$$\sigma^2 = \frac{(1-14.4)^2 + (2-14.4)^2 + (3-14.4)^2 + (4-14.4)^2 + (5-14.4)^2 + (6-14.4)^2 + (80-14.4)^2}{7}$$

$\sigma^2 = 719.10$



∴ we can conclude that when σ^2 is High then the spread will High.

ie

$\sigma^2 \propto$ spread of dataset.

② standard deviation (σ)

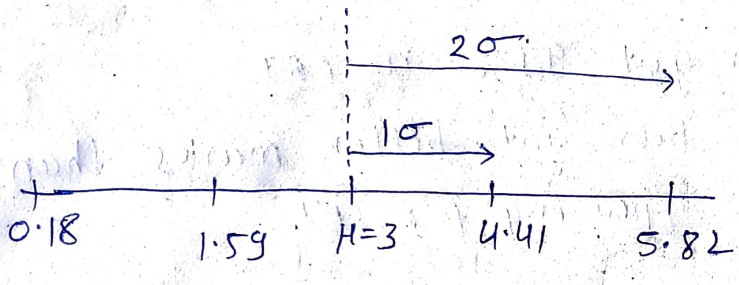
standard deviation measures the dispersion of a data set relative to its mean

σ = square root of variance. $= \sqrt{\sigma^2}$

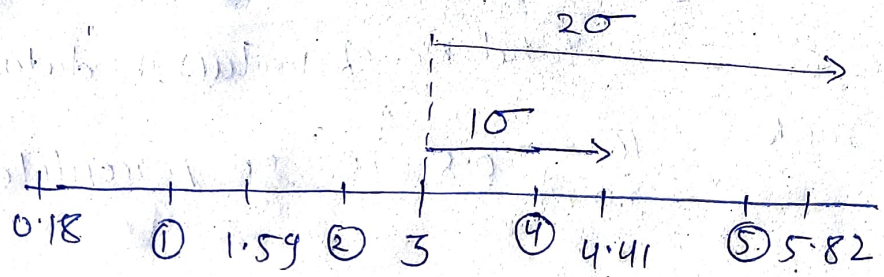
Eg: {1, 2, 3, 4, 5} $\mu = 3$

$\sigma^2 = 2$

$\sigma = \sqrt{2} = 1.41$



std deviation gives how many std. deviation away a number falls wrt mean.



4 falls between mean and 1σ.

5 falls between mean and 2σ.

① falls between mean and -2σ

2 falls between mean and -1σ.

④ percentile and Quartile.

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

$$\% \text{ of even Nos} = \frac{\text{count of even no.} \times 100}{\text{Total Numbers}} = \frac{4}{8} \times 100 = 50\%$$

percentile

A percentile is a value below which a certain percentage of observation lies.

Eg: if a person has got 99% in CAT this means person has got better marks than 99% of students appearing for CAT.

Dataset (must be sorted in Ascending order)

{2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

$$\text{percentile rank of } x = \frac{\text{No. of values below } x \text{ in dataset}}{\text{Total No. of values in dataset}}$$

$$\therefore \text{percentile rank of } 10 = \frac{16}{20} = 0.8 \text{ ie } 80 \text{ percentile}$$

$$\text{percentile rank of } 8 = \frac{9}{20} = 0.45 \text{ ie } 45 \text{ percentile}$$

$$\text{percentile rank of } 6 = \frac{7}{20} = 0.35 \text{ ie } 35 \text{ percentile}$$

$$\text{percentile rank of } 9 = \frac{14}{20} = 0.7 \text{ ie } 70 \text{ percentile}$$

* value at a given percentile.

$$\text{Value Index} = \frac{\text{Percentile}}{100} \times n+1$$ ← this gives index of element. (index starts from zero in dataset)

for previous example

$n = 20$, percentile $= 25$.

$$\text{value Index} = \frac{25}{100} \times (20+1)$$

$$= 5.25^{th} \text{ index.}$$

∴ the value will be avg of 5^{th} and 6^{th} index.

$$\text{value} = \frac{5+5}{2} = 5$$

$n = 20$, percentile $= 75$.

$$\text{value index} = \frac{75}{100} \times (20+1)$$

$$= 15.75^{th} \text{ index}$$

value = Avg of 15^{th} and 16^{th} index element

$$= \frac{9+10}{2}$$

value = 9.5

Here 25 percentile is first quartile

75 percentile is third quartile.

⑤ Five Number Summary.

It consists of Five different values wrt data set.

① minimum
② First quartile (25th percentile) Q_1
③ Median
④ Third quartile (75th percentile) Q_3.
⑤ maximum

This technique is used to remove outlies and Box plot is drawn to visualize.

Eg:

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

27 ↑ outlier

[Lower fence value ⟷ Higher fence value]

Lower fence = $Q_1 - 1.5\ IQR$

where Q_1 = 25th percentile

IQR = Inter quartile range = $Q_3 - Q_1$

const (1.5) is 1.5 std. deviation from mean

Higher fence = $Q_3 + 1.5(IQR)$

$Q_1 = \frac{25}{100} \times (20+1) = 5.25^{th}$ index

$\therefore Q_1 = \frac{3+3}{2} = (3)$

$Q_3 = \frac{75}{100} \times (20+1) = 15.75^{th}$ index

$Q_3 = \frac{7+8}{2} = (7.5)$

lower fence = $3 - 1.5(7.5-3) = -3.75$

Higher fence = $7.5 + 1.5(7.5-3) = 14.25$

∴ values should range from -3.75 to 14.25.

∴ outliers will be elements less than -3.75 and greater than 14.25

① minimum = 1

② $Q_1 = 3$

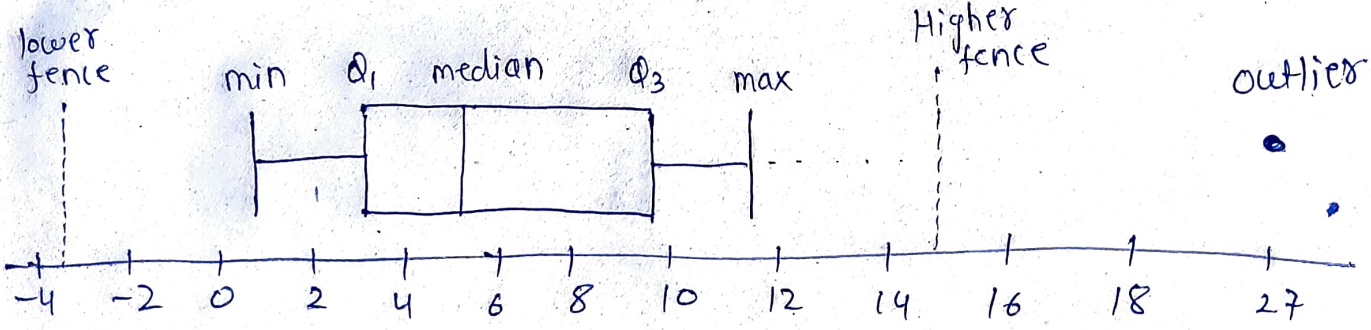
③ median = 5

④ $Q_3 = 7.5$

⑤ maximum = 9

} after removing outliers.

Box plot



Numbers below Higher fence and above lower fence will not be considered outliers.