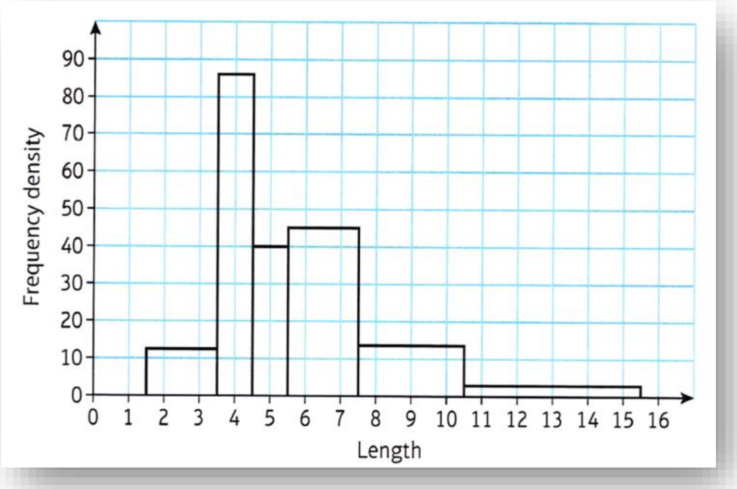
Histograms:

Grouped continuous data can be represented in a histogram. The area of each bar is proportional to the frequency of the class it represents

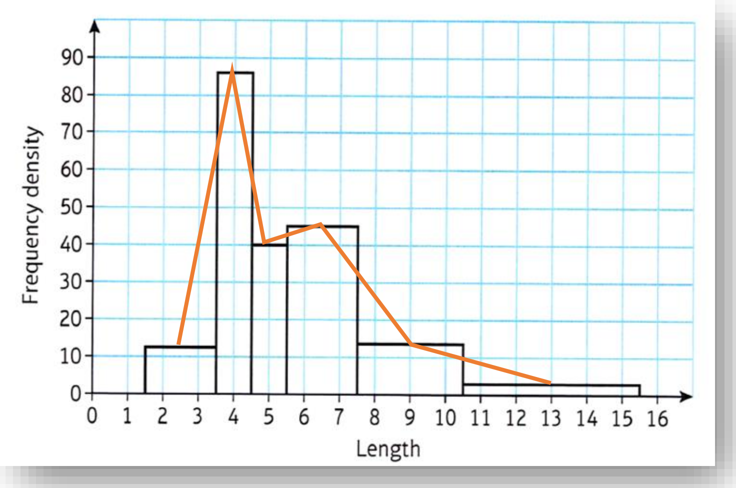


Notice that:

1. class widths don’t have to be the same
2. The height of each bar is not the frequency but frequency density.

$$Frequency\ density = \frac{frequency}{class\ width}$$

Frequency Polygon



We get a frequency polygon by Joining the mid points of the top of the bars of a histogram

Example 1:

In a random sample, 200 students were asked how long it took them to complete their homework the previous night. The times were recorded and summarised in the table below.

Time, t (minutes)	$25 \leqslant t < 30$	$30 \leqslant t < 35$	$35 \leqslant t < 40$	$40 \leqslant t < 50$	$50 \leqslant t < 80$
Frequency	55	39	68	32	6

a Draw a histogram and a frequency polygon to represent the data.

b Estimate how many students took between 36 and 45 minutes to complete their homework.

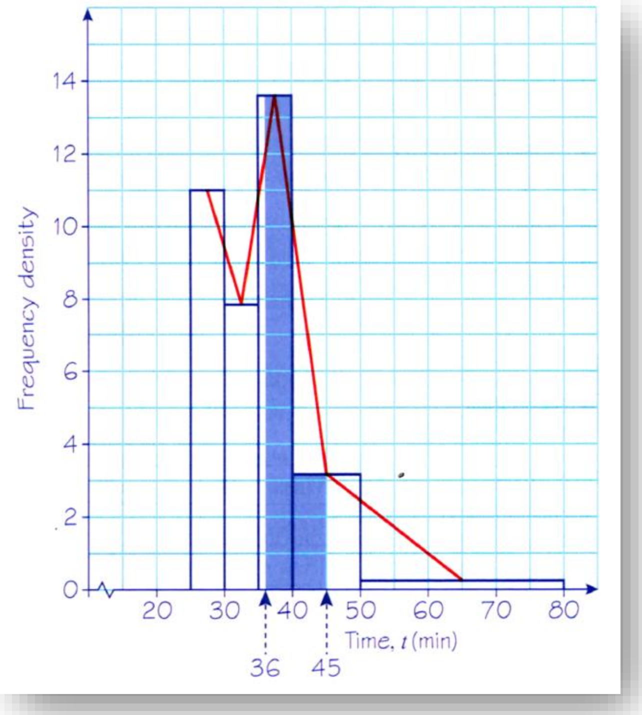
a.

Time (t)	f	$c.w$	$frequency\ density$
$25 \leq t < 30$	55	30 – 25 =5	$\frac{55}{5} = 11$
$30 \leq t < 35$	39	5	7.8
$35 \leq t < 40$	68	5	13.6
$40 \leq t < 50$	32	10	3.2
$50 \leq t < 80$	6	30	0.2

b. Number of students between 36 and 45 minutes is shown by the shaded area (f)

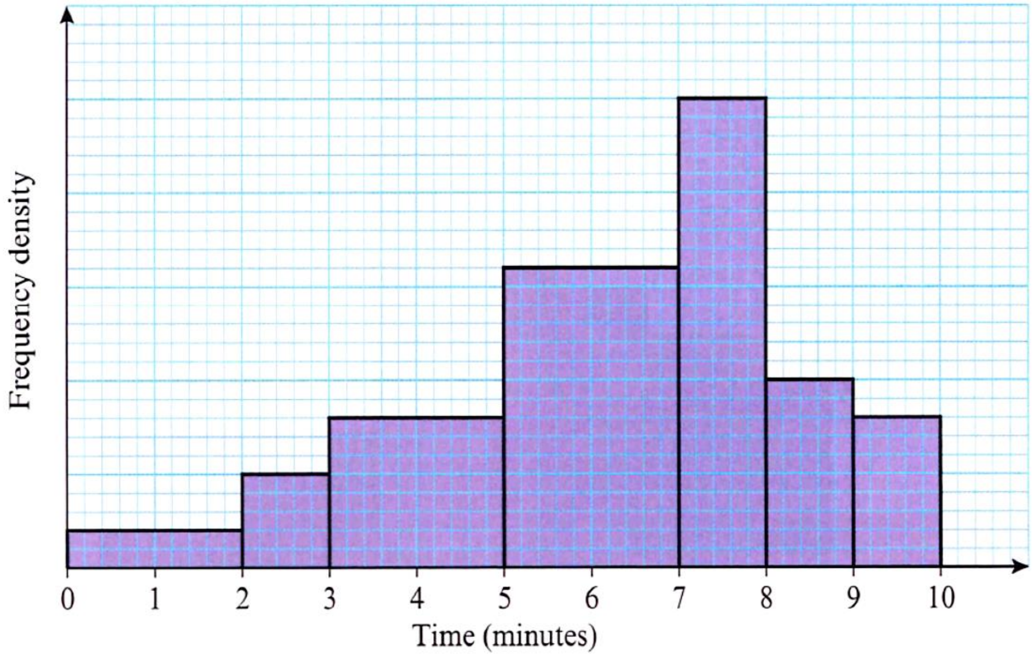
$(40 - 36) \times 13.6 + (45 - 40) \times 3.2$

$= 70.4$ (Since this is an estimate don't round the number)



Example 2:

The histogram below displays the information gathered from 100 people, regarding how long, in minutes, they took to complete a word puzzle.



a Why should a histogram be used to represent these data?

b Write down the underlying feature associated with each of the bars in a histogram.

c Given that 5 people completed the puzzle between 2 and 3 minutes, find the number of people who completed the puzzle between 0 and 2 minutes.

a. Time is continuous, and continuous data is well represented by a graph

b. The area of the bar is proportional to the frequency

c. Between 2 and 3 minutes there are 5x5 = 25 squares. So, 5 people are represented by 25 squares. $1\ square = \frac{5}{25} = \frac{1}{5}\ students$

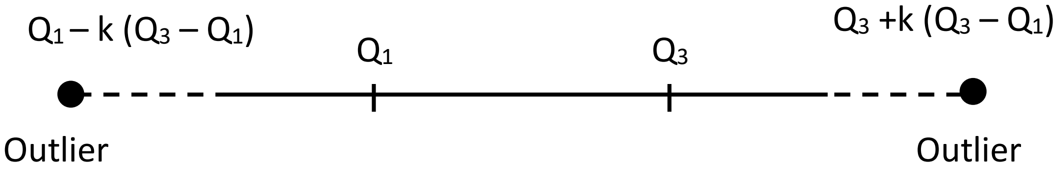
Between 0 and 2 we have 10x2 = 20 squares,

$20 \times \frac{1}{5} = 4\ persons$

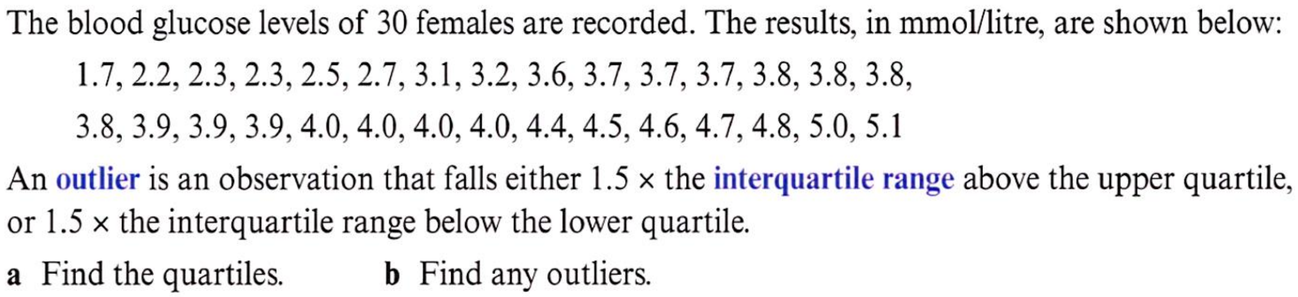
Outliers:

An outlier is an extreme value that lies outside the overall pattern.

So if we calculate $Q_3 - Q_1$ and consider this as the main body then $k(Q_3 - Q_1)$ is considered too far and an outlier in either direction. K will be given in the problem.



Example 1:


The blood glucose levels of 30 females are recorded. The results, in mmol/litre, are shown below:

1.7, 2.2, 2.3, 2.3, 2.5, 2.7, 3.1, 3.2, 3.6, 3.7, 3.7, 3.7, 3.8, 3.8, 3.8,
3.8, 3.9, 3.9, 3.9, 4.0, 4.0, 4.0, 4.0, 4.4, 4.5, 4.6, 4.7, 4.8, 5.0, 5.1

An **outlier** is an observation that falls either 1.5 × the **interquartile range** above the upper quartile, or 1.5 × the interquartile range below the lower quartile.

a Find the quartiles. **b** Find any outliers.

a. $k = 1.5$ $\quad$ Q_1 $\frac{30}{4} = 7.5$ so the 8th value = 3.2

Q_2 $\frac{30+1}{2} = 15.5$ so the 16th value = 3.8

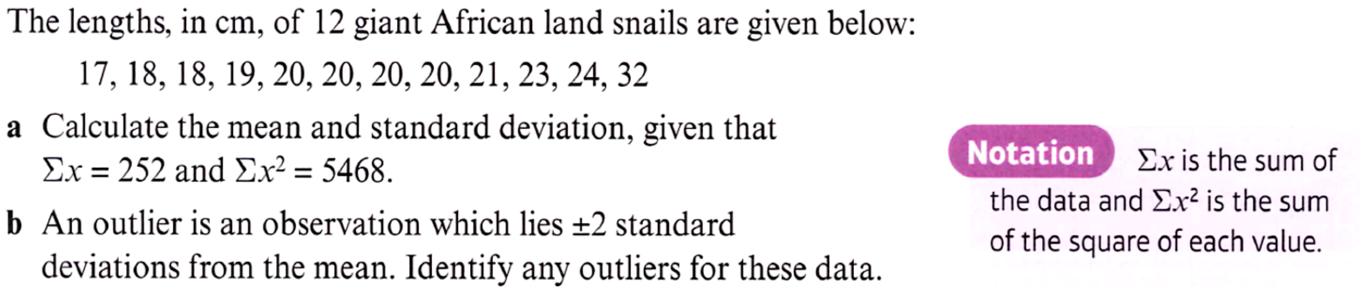
Q_3 $\frac{3}{4} \times 30 = 22.5$ so the 23rd value = 4.0

b. Outliers are less than $3.2 - 1.5(4 - 3.2) = 2$

or more than $4 + 1.5(4 - 3.2) = 5.2$

One value only 1.7 is considered an outlier

Example 2:



The lengths, in cm, of 12 giant African land snails are given below:

17, 18, 18, 19, 20, 20, 20, 20, 21, 23, 24, 32

a Calculate the mean and standard deviation, given that $\Sigma x = 252$ and $\Sigma x^2 = 5468$.

b An outlier is an observation which lies ±2 standard deviations from the mean. Identify any outliers for these data.

Notation Σx is the sum of the data and Σx^2 is the sum of the square of each value.

a. Mean $\bar{x} = \frac{\sum x}{n} = \frac{252}{12} = 21$ cm

Variance = $\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{5468}{12} - (21)^2 = 14.666$

Standard deviation = $\sqrt{14.666} = 3.83$

b. An outlier is any value less than mean – 2 x standard deviation

21 – 2(3.83) = 13.34 or more than mean + 2 x standard deviation

21 + 2(3.83) =28.66

32 is the only outlier

Anomalies:

Although an outlier is an odd value and unusual, yet it might be true for example in the last example a giant African snail that is 32 cm might not be the usual, yet might not be a wrong piece of data.

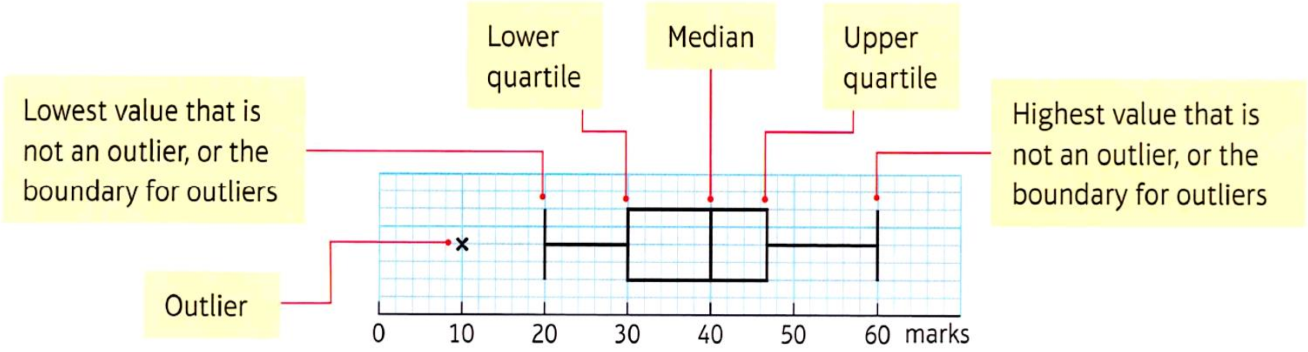
An anomaly is not just an outlier but an error that should be removed. This is called cleaning the data.

Example: If we collect the ages of people and find one entry that says 170 years!! That’s not just an outlier it’s a clear anomaly

Sometimes we use the symbol $\gg$ to say “much bigger than”

Box plots:

It shows many important features of the data as shown in the figure


Lower quartile
Median
Upper quartile
Lowest value that is not an outlier, or the boundary for outliers
Highest value that is not an outlier, or the boundary for outliers
Outlier
0
10
20
30
40
50
60
marks

Example:

The blood glucose levels of 30 males and females are recorded and the results are summarised as such:

Males

Lower quartile = 3.6
Upper quartile = 4.7
Median = 4.0
Lowest value = 1.4
Highest value = 5.2

Females

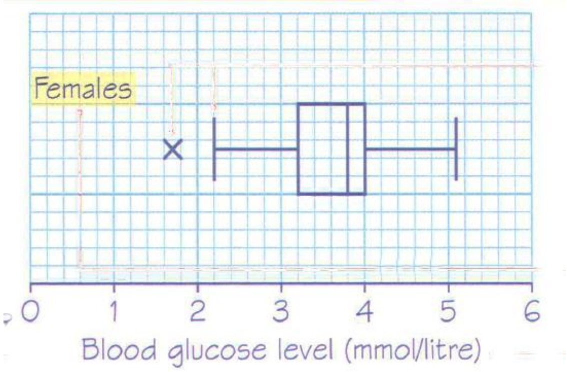
Lower quartile = 3.2
Upper quartile = 4.0
Median = 3.8
Outlier = 1.7
Lowest value = 2.2
Highest value = 5.1

a. Draw a box plot for the data for females.

An outlier is an observation that lies either 1.5 x the interquartile range above the upper quartile or 1.5 x the interquartile range below the lower quartile

b. Given that there is one outlier for males, use the same previous plot to add a box plot for males

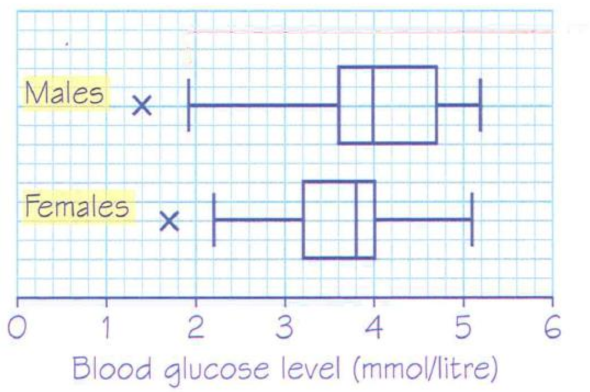
c. Compare the blood glucose levels between males and females a.



b. The interquartile is $Q_3 - Q_1 = 4.7 - 3.6 = 1.1$

Outliers are more than $4.7 + 1.5 \times 1.1 = 6.35$

or less than $3.6 - 1.5 \times 1.1 = 1.95$ so 1.4 is the only outlier



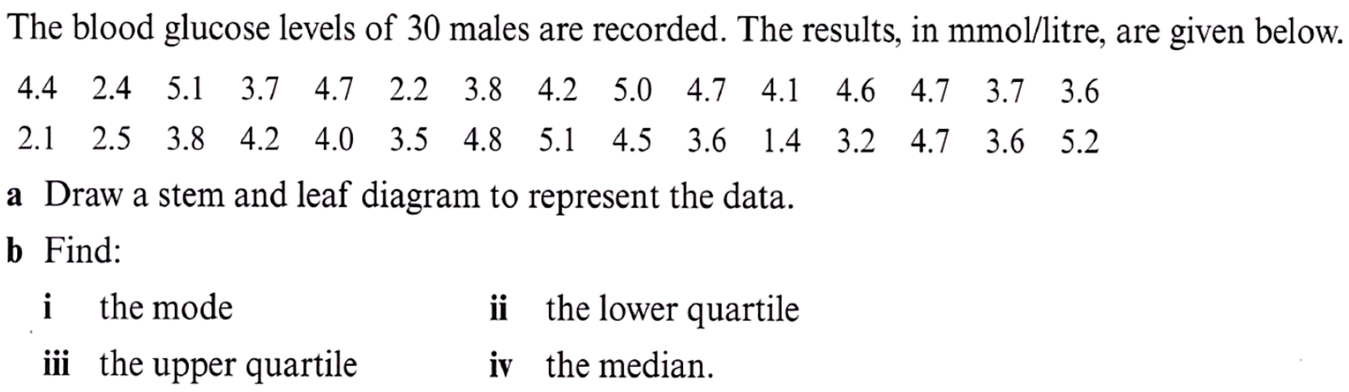
c. When comparing two sets of data you should talk about measure of location and measure of spread and then state your interpretation

So: The median blood glucose level of males is higher than that of females and the interquartile range in males in males is more than that of females.

Stem and leaf:

We have studied the “Stem and leaf” before so it’s better to work with examples

Example 1:



The blood glucose levels of 30 males are recorded. The results, in mmol/litre, are given below.

4.4	2.4	5.1	3.7	4.7	2.2	3.8	4.2	5.0	4.7	4.1	4.6	4.7	3.7	3.6
2.1	2.5	3.8	4.2	4.0	3.5	4.8	5.1	4.5	3.6	1.4	3.2	4.7	3.6	5.2

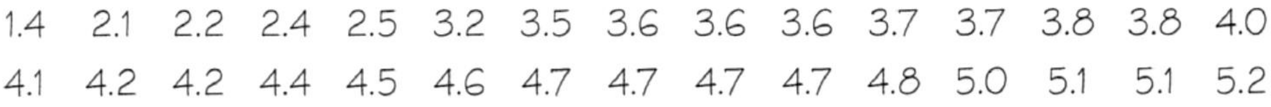
a Draw a stem and leaf diagram to represent the data.

b Find:

i the mode **ii** the lower quartile

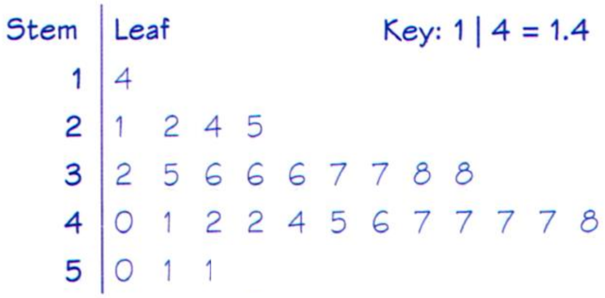
iii the upper quartile **iv** the median.

First, we have to rearrange the data



1.4	2.1	2.2	2.4	2.5	3.2	3.5	3.6	3.6	3.6	3.7	3.7	3.8	3.8	4.0
4.1	4.2	4.2	4.4	4.5	4.6	4.7	4.7	4.7	4.7	4.8	5.0	5.1	5.1	5.2

Second, we start to choose a stem and start drawing, making sure to include a key to show how to read the figure



Key: 1 | 4 = 1.4

Stem	Leaf
1	4
2	1 2 4 5
3	2 5 6 6 6 7 7 8 8
4	0 1 2 2 4 5 6 7 7 7 7 8
5	0 1 1

i. The mode is 4.7

ii. The lower quartile:

$\frac{30}{4}$ = 7.5 so we take

the 8th term = 3.6

iii. The upper quartile:

$\frac{3}{4} \times 30$ = 22.5 so we take the 23rd term = 4.7

iv. The median: $\frac{30+1}{2}$ = 15.5 so the average of the 15th and 16th = 4.05

Exercise 2: (Back to Back stem and leaf)

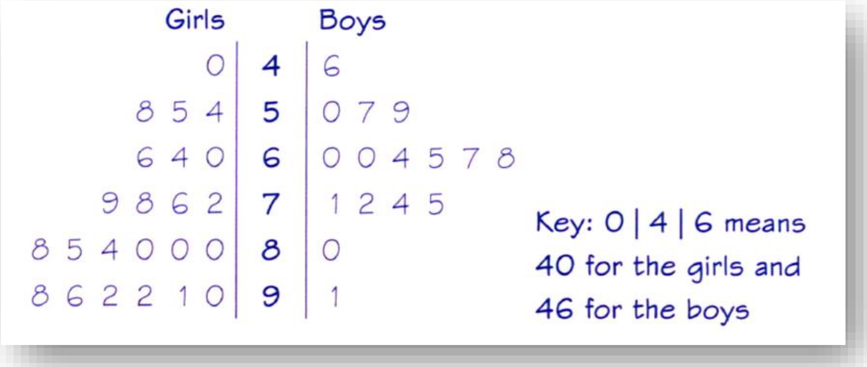
Achara recorded the resting pulse rate for the 16 boys and 23 girls in her year at school. The results were as follows:

Girls						Boys			
55	80	84	91	80	92	80	60	91	65
98	40	60	64	66	72	67	59	75	46
96	85	88	90	76	54	72	71	74	57
58	92	78	80	79		64	60	50	68

a Construct a back to back stem and leaf diagram to represent these data.

b Comment on your results.

a.



Girls		Boys
0	4	6
8 5 4	5	0 7 9
6 4 0	6	0 0 4 5 7 8
9 8 6 2	7	1 2 4 5
8 5 4 0 0 0	8	0
8 6 2 2 1 0	9	1

Key: 0 | 4 | 6 means 40 for the girls and 46 for the boys

N.B: In a back to back figure we read in both directions

0|4|6

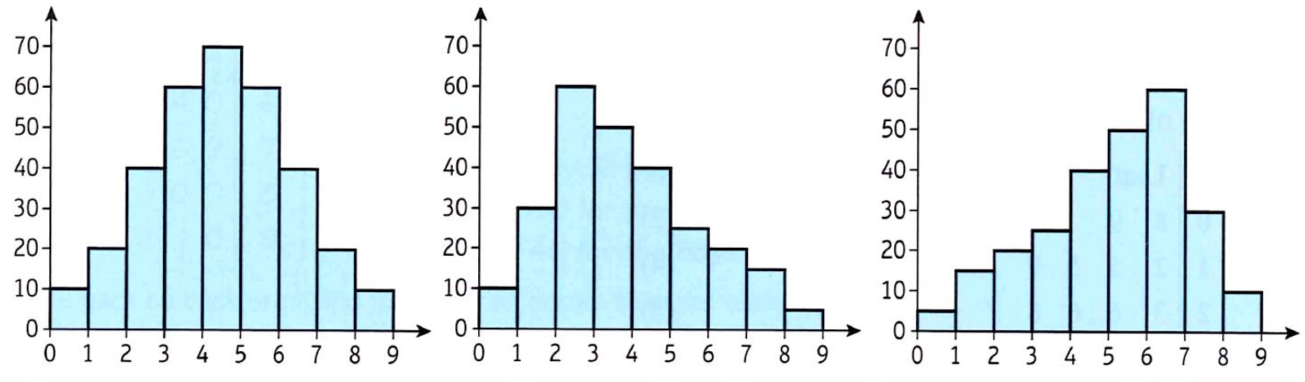
For girls ← 40

For boys → 46

b. The resting pulse of boys is lower than that of girls

Skewness:

■ A distribution can be symmetrical, have a positive skew or have a negative skew.



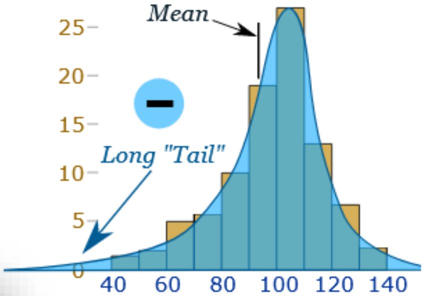
This distribution is said to be symmetrical

This distribution is said to have a positive skew

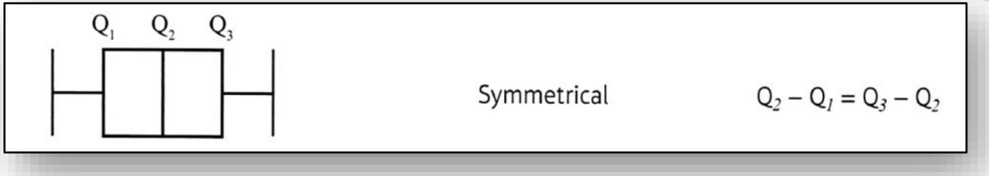
This distribution is said to have a negative skew

When the tail goes to the left (The negative side)

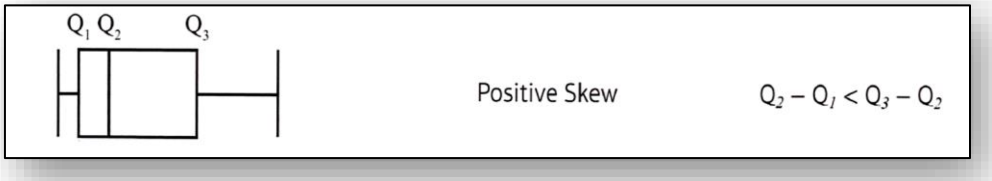
We call it a negative skew.



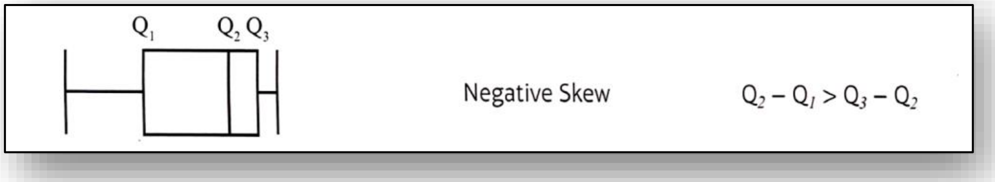
Using a box plot:



Symmetrical $Q_2 - Q_1 = Q_3 - Q_2$

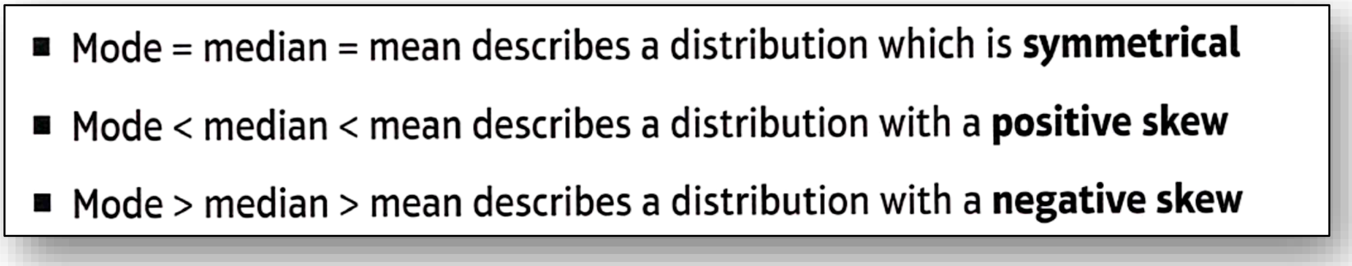


Positive Skew $Q_2 - Q_1 < Q_3 - Q_2$



Negative Skew $Q_2 - Q_1 > Q_3 - Q_2$

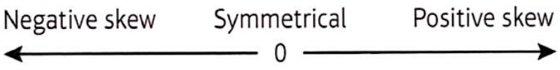
Using Mode, Median and Mean



- Mode = median = mean describes a distribution which is **symmetrical**
- Mode < median < mean describes a distribution with a **positive skew**
- Mode > median > mean describes a distribution with a **negative skew**

Using the skewness formula

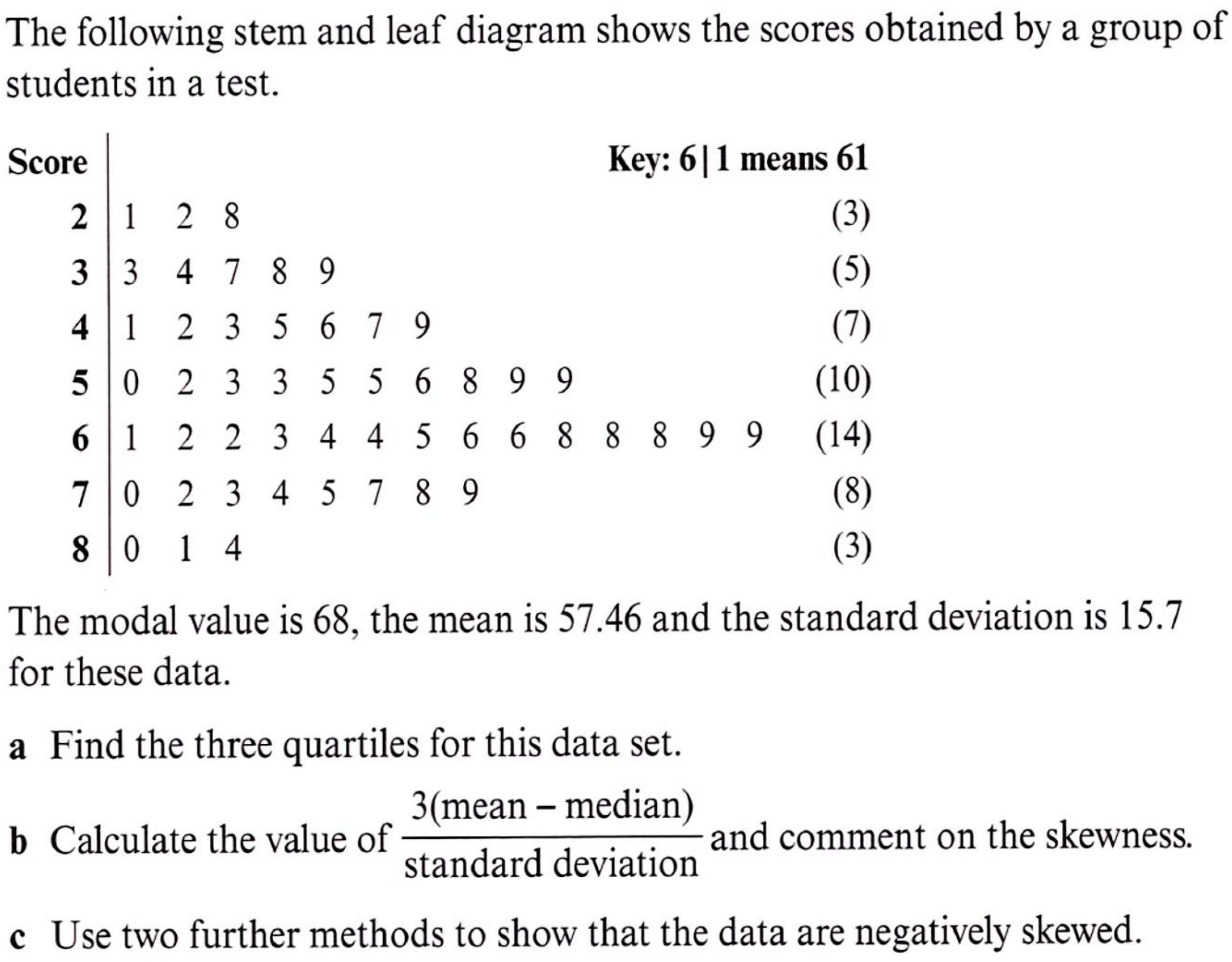
You can also calculate $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$ which tells you how **skewed** the data are.



- A value of 0 implies that the mean = median and the distribution is **symmetrical**
- A positive value implies that the median < mean and the distribution is **positively skewed**
- A negative value implies that median > mean and the distribution is **negatively skewed**

The further from 0 the value is, the more likely the data will be skewed.

Example:



The following stem and leaf diagram shows the scores obtained by a group of students in a test.

Score		Key: 6|1 means 61
2	1 2 8	(3)
3	3 4 7 8 9	(5)
4	1 2 3 5 6 7 9	(7)
5	0 2 3 3 5 5 6 8 9 9	(10)
6	1 2 2 3 4 4 5 6 6 8 8 8 9 9	(14)
7	0 2 3 4 5 7 8 9	(8)
8	0 1 4	(3)

The modal value is 68, the mean is 57.46 and the standard deviation is 15.7 for these data.

a Find the three quartiles for this data set.

b Calculate the value of $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$ and comment on the skewness.

c Use two further methods to show that the data are negatively skewed.

Total number of students = 50

a. Q_1: $\frac{50}{4}$ =12.5 so we take the 13th value = 46

Q_2: $\frac{50+1}{2}$ =25.5 so we take the mean of the 25th and 26th values = 60

Q_3: $\frac{3\times 50}{4}$ =37.5 so we take the 38th value = 69

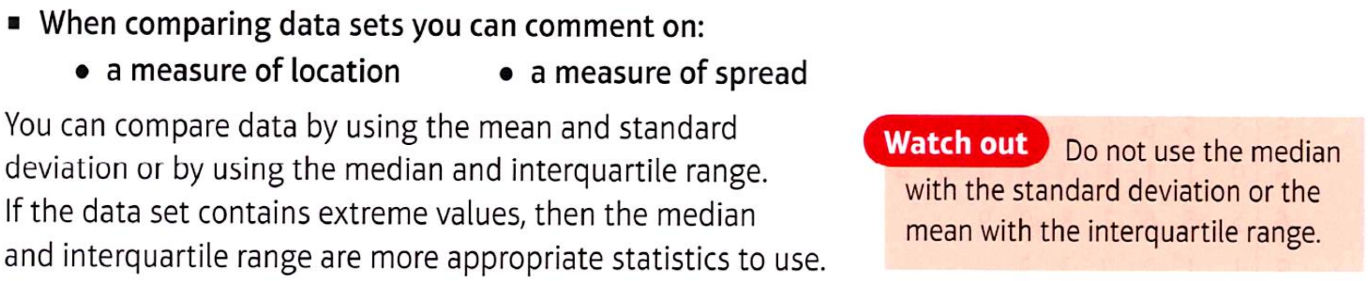
b. $\frac{3(57.46-60)}{15.7}$ = - 0.485 The data is negatively skewed

c. i. $Q_2 - Q_1 = 60 - 46 = 14$, $Q_3 - Q_2 = 69 - 60 = 9$

$Q_2 - Q_1 > Q_3 - Q_2$ The data is negatively skewed

ii. Median (60) > Mean (57.46) The data is negatively skewed

Comparing data:



- **When comparing data sets you can comment on:**
  - **a measure of location**
  - **a measure of spread**

You can compare data by using the mean and standard deviation or by using the median and interquartile range. If the data set contains extreme values, then the median and interquartile range are more appropriate statistics to use.

Watch out Do not use the median with the standard deviation or the mean with the interquartile range.