CHAPTER 11

Statistics

- Ogives are not always grounded at $(0; 0)$.
- Histograms must have bars of equal widths.
- The formula for population variance is used and not sample variance.
- Encourage learners to use the STATS functions on their calculators.
- Learners do not need to draw scatter plots, they need only identify outliers.
- Discuss the misuse of statistics in the real world and encourage awareness.

11.1 Revision

Exercise 11 – 1: Revision

1. For each of the following data sets, compute the mean and all the quartiles. Round your answers to one decimal place.

   a) $-3{,}4 ; -3{,}1 ; -6{,}1 ; -1{,}5 ; -7{,}8 ; -3{,}4 ; -2{,}7 ; -6{,}2$

   b) $-6 ; -99 ; 90 ; 81 ; 13 ; -85 ; -60 ; 65 ; -49$

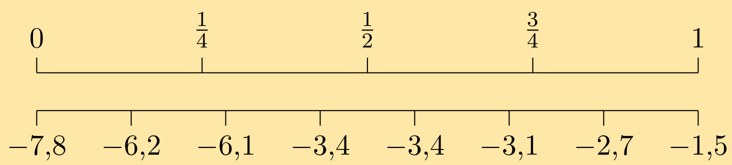
   c) $7 ; 45 ; 11 ; 3 ; 9 ; 35 ; 31 ; 7 ; 16 ; 40 ; 12 ; 6$

   Solution:

   a) Mean:

   $$\begin{aligned}\overline{x} &= \frac{(-3{,}4) + (-3{,}1) + (-6{,}1) + (-1{,}5) + (-7{,}8) + (-3{,}4) + (-2{,}7) + (-6{,}2)}{8} \\ &\approx -4{,}3\end{aligned}$$

   To compute the quartiles, we order the data:
   $-7{,}8 ; -6{,}2 ; -6{,}1 ; -3{,}4 ; -3{,}4 ; -3{,}1 ; -2{,}7 ; -1{,}5$
   We use the diagram below to find at or between which values the quartiles lie.

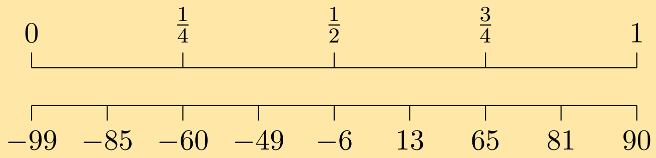
   

   For the first quartile the position is between the second and third values. The second value is $-6{,}2$ and the third value is $-6{,}1$, which means that the first quartile is $\frac{-6{,}2-6{,}1}{2} = -6{,}15$.
   For the median (second quartile) the position is halfway between the fourth and fifth values. Since both these values are $-3{,}4$, the median is $-3{,}4$.
   For the third quartile the position is between the sixth and seventh values. Therefore the third quartile is $-2{,}9$.

b) Mean:

$$\overline{x} = \frac{(-6)+(-99)+(90)+(81)+(13)+(-85)+(-60)+(65)+(-49)}{9}$$
$$\approx -5{,}6$$

To compute the quartiles, we order the data:
-99 ; -85 ; -60 ; -49 ; -6 ; 13 ; 65 ; 81 ; 90
We use the diagram below to find at or between which values the quartiles lie.

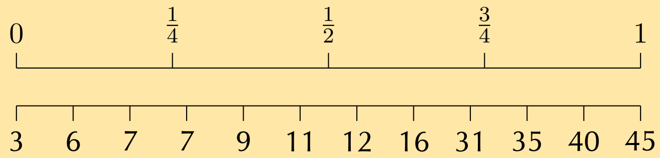


We see that the quartiles are at -60; -6; and 65.

c) The mean is $\overline{x} = 18{,}5$.
To compute the quartiles, we order the data:

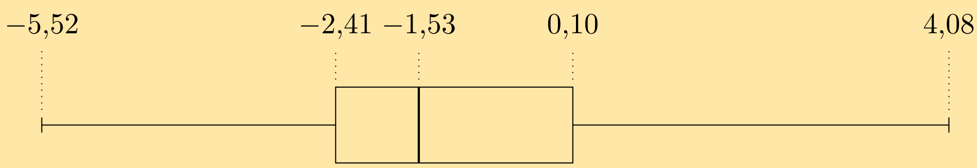
$$3\ ;\ 6\ ;\ 7\ ;\ 7\ ;\ 9\ ;\ 11\ ;\ 12\ ;\ 16\ ;\ 31\ ;\ 35\ ;\ 40\ ;\ 45$$

We use the diagram below to find at or between which values the quartiles lie.



For the first quartile the position is between the third and fourth values. Since both these values are equal to 7, the first quartile is 7.
For the median (second quartile) the position is halfway between the sixth and seventh values. The sixth value is 11 and the seventh value is 12, which means that the median is $\frac{11+12}{2} = 11{,}5$.
For the third quartile the position is between the ninth and tenth values. Therefore the third quartile is $\frac{31+35}{2} = 33$.

2. Use the following box and whisker diagram to determine the range and inter-quartile range of the data.



Solution:

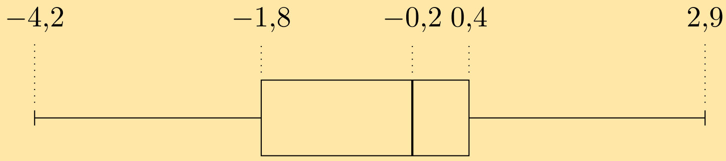
The range is the difference between the minimum and maximum values. From the box-and-whisker diagram, the minimum is $-5{,}52$ and the maximum is $4{,}08$. Therefore the range is $4{,}08 - (-5{,}52) = 9{,}6$.

The inter-quartile range is the difference between the first and third quartiles. From the box-and-whisker diagram, the first quartile is $-2{,}41$ and the third quartile is $0{,}10$. Therefore the range is $0{,}10 - (-2{,}41) = 2{,}51$.

3. Draw the box and whisker diagram for the following data.

$$0{,}2\ ;\ -0{,}2\ ;\ -2{,}7\ ;\ 2{,}9\ ;\ -0{,}2\ ;\ -4{,}2\ ;\ -1{,}8\ ;\ 0{,}4\ ;\ -1{,}7\ ;\ -2{,}5\ ;\ 2{,}7\ ;\ 0{,}8\ ;\ -0{,}5$$

Solution:

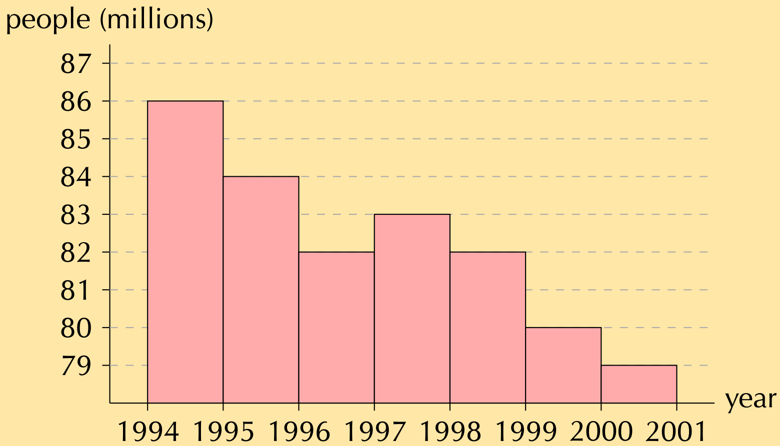


11.2 Histograms

Frequency polygons

Exercise 11 – 2: Histograms

1. Use the histogram below to answer the following questions. The histogram shows the number of people born around the world each year. The ticks on the x-axis are located at the start of each year.



a) How many people were born between the beginning of 1994 and the beginning of 1996?

b) Is the number people in the world population increasing or decreasing? (Ignore the rate at which people are dying for this question.)

c) How many more people were born in 1994 than in 1997?

Solution:

a) $86 + 84 = 170$ million

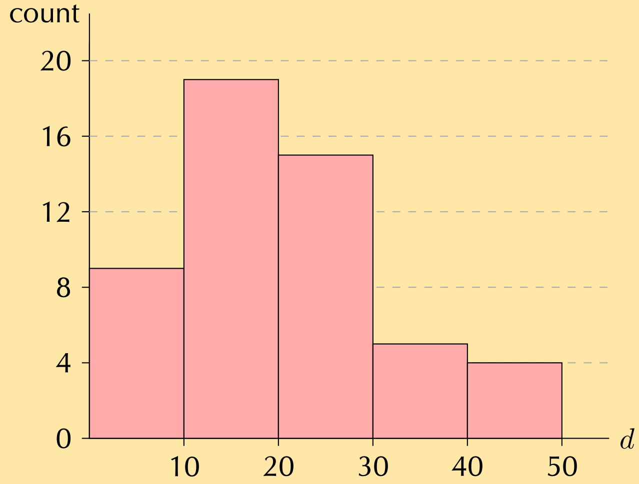
b) Even though the rate at which people are born seems to be decreasing, there are still new people born every year and so the world population is increasing.

c) $86 - 83 = 3$ million

2. In a traffic survey, a random sample of 50 motorists were asked the distance they drove to work daily. The results of the survey are shown in the table below. Draw a histogram to represent the data.

d (km)	$0 < d \leq 10$	$10 < d \leq 20$	$20 < d \leq 30$	$30 < d \leq 40$	$40 < d \leq 50$
f	9	19	15	5	4

Solution:

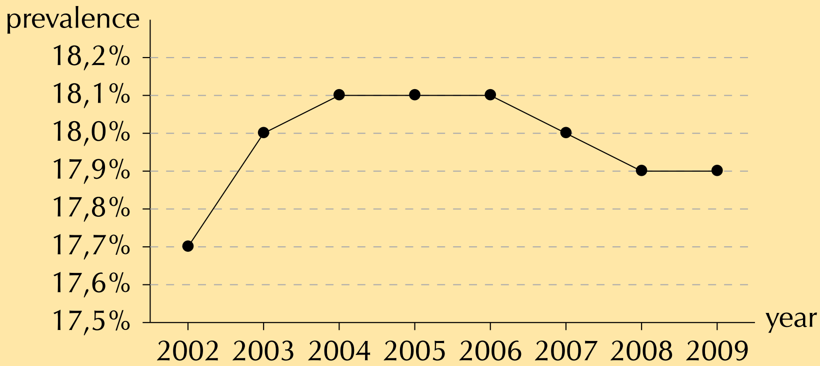


3. Below is data for the prevalence of HIV in South Africa. HIV prevalence refers to the percentage of people between the ages of 15 and 49 who are infected with HIV.

year	2002	2003	2004	2005	2006	2007	2008	2009
prevalence (%)	17,7	18,0	18,1	18,1	18,1	18,0	17,9	17,9

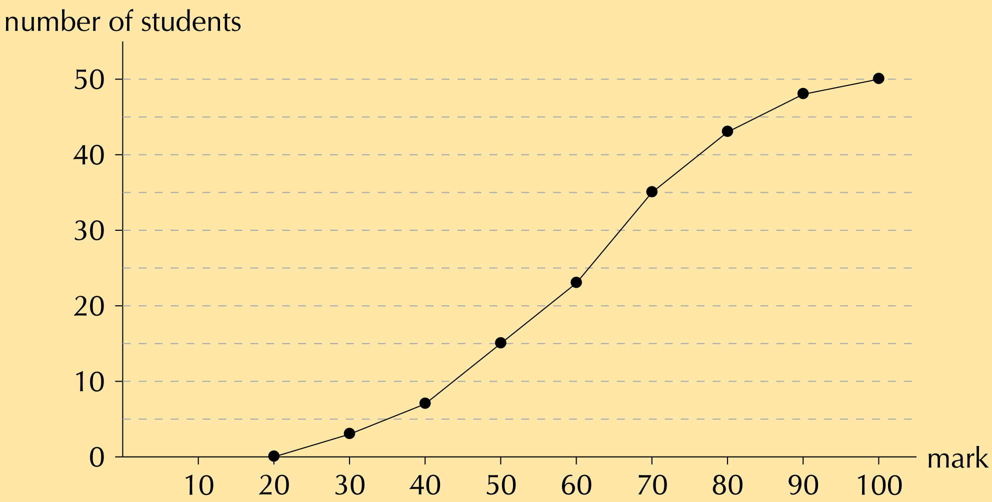
Draw a frequency polygon of this data set.

Solution:



Exercise 11 – 3: Ogives

1. Use the ogive to answer the questions below. Marks give as a percentage (%).



a) How many students got between 50% and 70%?

b) How many students got at least 70%?

c) Compute the average mark for this class, rounded to the nearest integer.

Solution:

a) The cumulative plot shows that 15 students got below 50% and 35 students got below 70%. Therefore $35 - 15 = 20$ students got between 50% and 70%.

b) The cumulative plot shows that 35 students got below 70% and that there are 50 students in total. Therefore $50 - 35 = 15$ students got at least (greater than or equal to) 70%.

c) To compute the average, we first need to use the ogive to determine the frequency of each interval. The frequency of an interval is the difference between the cumulative counts at the top and bottom of the interval on the ogive. It might be difficult to read the exact cumulative count for some of the points on the ogive. But since the final answer will be rounded to the nearest integer, small errors in the counts will not make a difference. The table below summarises the counts.

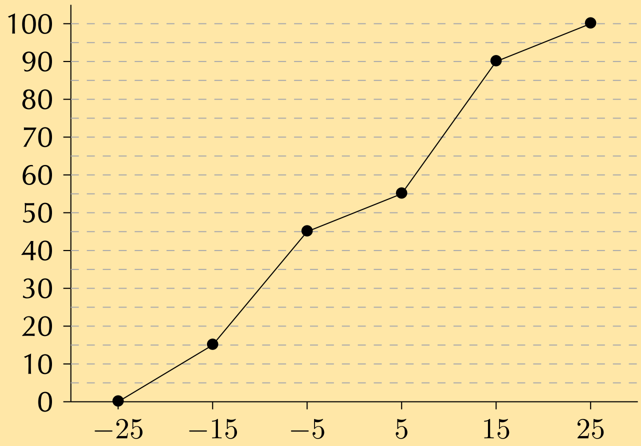
Interval	$[20, 30)$	$[30, 40)$	$[40, 50)$	$[50, 60)$
f	3	4	8	8
Interval	$[60, 70)$	$[70, 80)$	$[80, 90)$	$[90, 100)$
f	12	8	5	2

The average is then the centre of each interval, weighted by the count in that interval.

$$\frac{3 \times 25 + 4 \times 35 + 8 \times 45 + 8 \times 55 + 12 \times 65 + 8 \times 75 + 5 \times 85 + 2 \times 95}{3 + 4 + 8 + 8 + 12 + 8 + 5 + 2}$$
$$= 60{,}2$$

The average mark, rounded to the nearest integer, is 60%.

2. Draw the histogram corresponding to this ogive.



Solution:

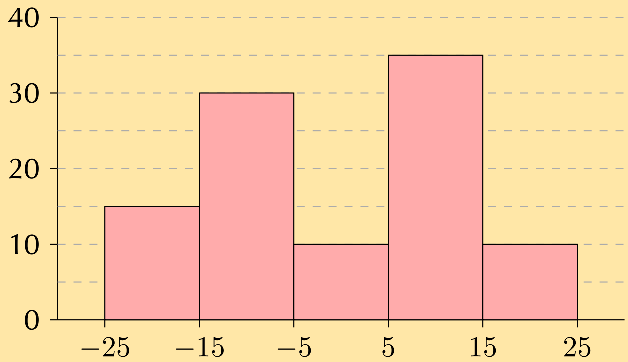
To draw the histogram we need to determine the count in each interval.

Firstly, we can find the intervals by looking where the points are plotted on the ogive. Since the points are at x-coordinates of -25; -15; -5; 5; 15 and 25, it means that the intervals are $[-25;-15)$, etc.

To get the count in each interval we subtract the cumulative count at the start of the interval from the cumulative count at the end of the interval.

Interval	$[-25;-15)$	$[-15;-5)$	$[-5;5)$	$[5;15)$	$[15;25)$
Count	15	30	10	35	10

From these counts we can draw the following histogram:



3. The following data set lists the ages of 24 people.

   $2; 5; 1; 76; 34; 23; 65; 22; 63; 45; 53; 38$

   $4; 28; 5; 73; 79; 17; 15; 5; 34; 37; 45; 56$

   Use the data to answer the following questions.

   a) Using an interval width of 8 construct a cumulative frequency plot.
   b) How many are below 30?
   c) How many are below 60?
   d) Giving an explanation, state below what value the bottom 50% of the ages fall.

e) Below what value do the bottom 40% fall?

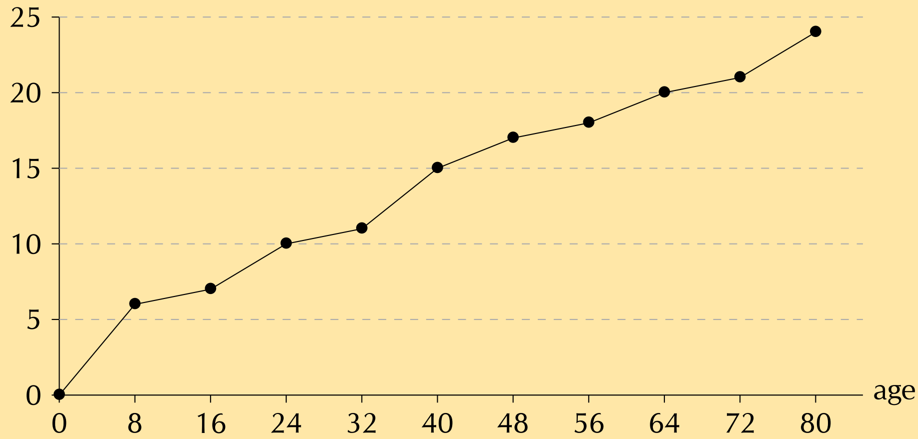
f) Construct a frequency polygon.

Solution:

a) The table below shows the number of people in each age bracket of width 8.

Interval	$[0; 8)$	$[8; 16)$	$[16; 24)$	$[24; 32)$	$[32; 40)$
Count	6	1	3	1	4
Cumulative	6	7	10	11	15
Interval	$[40; 48)$	$[48; 56)$	$[56; 64)$	$[64; 72)$	$[72; 80)$
Count	2	1	2	1	3
Cumulative	17	18	20	21	24

From this table we can draw the cumulative frequency plot:



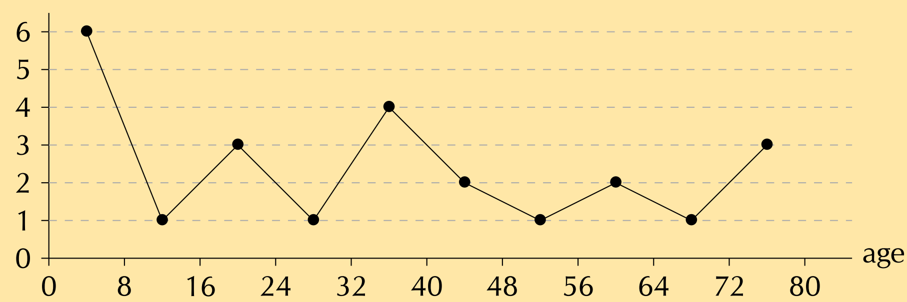
b) 11 people

c) 19 people

d) This question is asking for the median of the data set. The median is, by definition, the value below which 50% of the data lie. Since there are 24 values, the median lies between the middle two values, giving 34.

e) There are 24 values. By drawing a number line, as we do for determining quartiles, we can see that the 40% point is between the tenth and eleventh values. The tenth value is 23 and the eleventh value is 28. Therefore 40% of the values lie below $\frac{23 + 28}{2} = 25{,}5$.

f) We already have all the values needed to construct the frequency polygon in the table of values above.



4. The weights of bags of sand in grams is given below (rounded to the nearest tenth):

   $50{,}1$; $40{,}4$; $48{,}5$; $29{,}4$; $50{,}2$; $55{,}3$; $58{,}1$; $35{,}3$; $54{,}2$; $43{,}5$

   $60{,}1$; $43{,}9$; $45{,}3$; $49{,}2$; $36{,}6$; $31{,}5$; $63{,}1$; $49{,}3$; $43{,}4$; $54{,}1$

   a) Decide on an interval width and state what you observe about your choice.
   b) Give your lowest interval.
   c) Give your highest interval.
   d) Construct a cumulative frequency graph and a frequency polygon.
   e) Below what value do 53% of the cases fall?
   f) Below what value of 60% of the cases fall?

   Solution:

   a) Learner-dependent answer.
   b) Learner-dependent answer.
   c) Learner-dependent answer.
   d) Learner-dependent answer.
   e) $49{,}25$
   f) $49{,}7$

11.4 Variance and standard deviation

Exercise 11 – 4: Variance and standard deviation

1. Bridget surveyed the price of petrol at petrol stations in Cape Town and Durban. The data, in rands per litre, are given below.

Cape Town	$3{,}96$	$3{,}76$	$4{,}00$	$3{,}91$	$3{,}69$	$3{,}72$
Durban	$3{,}97$	$3{,}81$	$3{,}52$	$4{,}08$	$3{,}88$	$3{,}68$

   a) Find the mean price in each city and then state which city has the lower mean.
   b) Find the standard deviation of each city's prices.
   c) Which city has the more consistently priced petrol? Give reasons for your answer.

   Solution:

   a) Cape Town: $3{,}84$. Durban: $3{,}82$. Durban has the lower mean.
   b) Standard deviation:

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

For Cape Town:

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum (x - (3{,}84))^2}{6}} \\ &= \sqrt{\frac{0{,}0882}{6}} \\ &= \sqrt{0{,}0147} \\ &\approx 0{,}121\end{aligned}$$

For Durban:

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum \left(x - (3{,}82\dot{3})\right)^2}{6}} \\ &= \sqrt{\frac{0{,}20\dot{3}}{6}} \\ &= \sqrt{0{,}033\dot{8}} \\ &\approx 0{,}184\end{aligned}$$

c) The standard deviation of Cape Town's prices is lower than that of Durban's. That means that Cape Town has more consistent (less variable) prices than Durban.

2. Compute the mean and variance of the following set of values.

   $150 ; 300 ; 250 ; 270 ; 130 ; 80 ; 700 ; 500 ; 200 ; 220 ; 110 ; 320 ; 420 ; 140$

   Solution:

   Mean = $270{,}7$. Variance = $27\ 435{,}2$.

3. Compute the mean and variance of the following set of values.

   $-6{,}9 ; -17{,}3 ; 18{,}1 ; 1{,}5 ; 8{,}1 ; 9{,}6 ; -13{,}1 ; -14{,}0 ; 10{,}5 ; -14{,}8 ; -6{,}5 ; 1{,}4$

   Solution:

   Mean = $-1{,}95$. Variance = $127{,}5$.

4. The times for 8 athletes who ran a 100 m sprint on the same track are shown below. All times are in seconds.

   $10{,}2 ; 10{,}8 ; 10{,}9 ; 10{,}3 ; 10{,}2 ; 10{,}4 ; 10{,}1 ; 10{,}4$

   a) Calculate the mean time.
   b) Calculate the standard deviation for the data.
   c) How many of the athletes' times are more than one standard deviation away from the mean?

   Solution:

   a) $\overline{x} = 10{,}4$
   b) $\sigma = 0{,}27$
   c) The mean is $10{,}4$ and the standard deviation is $0{,}27$. Therefore the interval containing all values that are one standard deviation from the mean is $[10{,}4 - 0{,}27; 10{,}4 + 0{,}27] = [10{,}13; 10{,}67]$. We are asked how many values are **further** than one standard deviation from the mean, meaning **outside** the interval. There are 3 values from the data set outside the interval.

5. The following data set has a mean of $14{,}7$ and a variance of $10{,}01$.

$$18 \ ; \ 11 \ ; \ 12 \ ; \ a \ ; \ 16 \ ; \ 11 \ ; \ 19 \ ; \ 14 \ ; \ b \ ; \ 13$$

Compute the values of a and b.

Solution:

From the formula of the mean we have

$$\begin{aligned} 14{,}7 &= \frac{114 + a + b}{10} \\ \therefore a + b &= 147 - 114 \\ \therefore a &= 33 - b \end{aligned}$$

From the formula of the variance we have

$$\begin{aligned} \sigma^2 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \\ \therefore 10{,}01 &= \frac{69{,}12 + (a - 14{,}7)^2 + (b - 14{,}7)^2}{10} \end{aligned}$$

Substitute $a = 33 - b$ into this equation to get

$$\begin{aligned} 10{,}01 &= \frac{69{,}12 + (18{,}3 - b)^2 + (b - 14{,}7)^2}{10} \\ \therefore 100{,}1 &= 2b^2 - 66b + 620{,}1 \\ \therefore 0 &= b^2 - 33b + 260 \\ &= (b - 13)(b - 20) \end{aligned}$$

Therefore $b = 13$ or $b = 20$.

Since $a = 33 - b$ we have $a = 20$ or $a = 13$. So, the two unknown values in the data set are 13 and 20.

We do not know which of these is a and which is b since the mean and variance tell us nothing about the order of the data.

11.5 Symmetric and skewed data

Exercise 11 – 5: Symmetric and skewed data

1. Is the following data set symmetric, skewed right or skewed left? Motivate your answer.

   $27 \ ; 28 \ ; 30 \ ; 32 \ ; 34 \ ; 38 \ ; 41 \ ; 42 \ ; 43 \ ; 44 \ ; 46 \ ; 53 \ ; 56 \ ; 62$

   Solution:

   The statistics of the data set are

   - mean: $41{,}1$;
   - first quartile: 33;

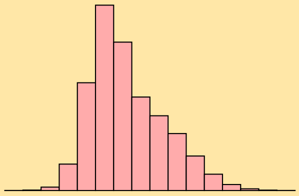
- median: $41{,}5$;
- third quartile: 45.

We can conclude that the data set is skewed left for two reasons.

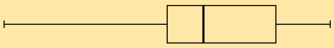
- The mean is less than the median. There is only a very small difference between the mean and median, so this is not a very strong reason.
- A better reason is that the median is closer to the third quartile than the first quartile.

2. State whether each of the following data sets are symmetric, skewed right or skewed left.

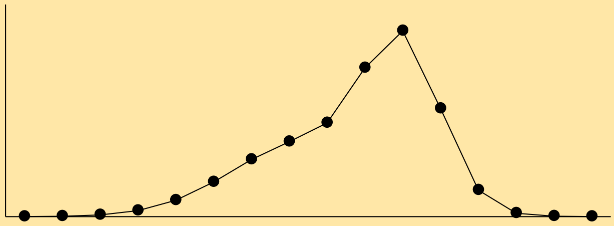
   a) A data set with this histogram:



   b) A data set with this box and whisker plot:



   c) A data set with this frequency polygon:



   d) The following data set:
   $11{,}2$; 5 ; $9{,}4$; $14{,}9$; $4{,}4$; $18{,}8$; $-0{,}4$; $10{,}5$; $8{,}3$; $17{,}8$

Solution:

a) skewed right

b) skewed right

c) skewed left

d) The statistics of the data set are

- mean: $9{,}99$;
- first quartile: $6{,}65$;
- median: $9{,}95$;
- third quartile: $13{,}05$.

Note that we get contradicting indications from the different ways of determining whether the data is skewed right or left.

- The mean is slightly greater than the median. This would indicate that the data set is skewed right.
- The median is slightly closer to the third quartile than the first quartile. This would indicate that the data set is skewed left.

Since these differences are so small and since they contradict each other, we conclude that the data set is symmetric.

3. Two data sets have the same range and interquartile range, but one is skewed right and the other is skewed left. Sketch the box and whisker plot for each of these data sets. Then, invent data (6 points in each data set) that matches the descriptions of the two data sets.

   Solution:

   Learner-dependent answer.

11.6 Identification of outliers

Exercise 11 – 6: Outliers

1. For each of the following data sets, draw a box and whisker diagram and determine whether there are any outliers in the data.

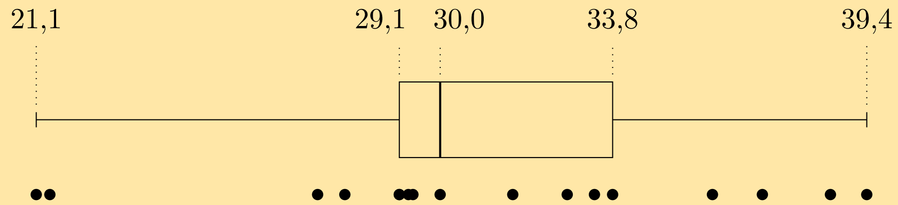
   a) $30\ ;\ 21{,}4\ ;\ 39{,}4\ ;\ 33{,}4\ ;\ 21{,}1\ ;\ 29{,}3\ ;\ 32{,}8\ ;\ 31{,}6\ ;\ 36\ ;$
   $27{,}9\ ;\ 27{,}3\ ;\ 29{,}4\ ;\ 29{,}1\ ;\ 38{,}6\ ;\ 33{,}8\ ;\ 29{,}1\ ;\ 37{,}1$

   b) $198\ ;\ 166\ ;\ 175\ ;\ 147\ ;\ 125\ ;\ 194\ ;\ 119\ ;\ 170\ ;\ 142\ ;\ 148$

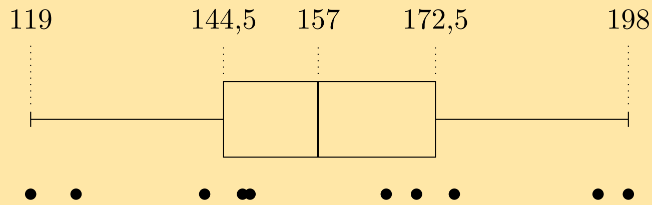
   c) $7{,}1\ ;\ 9{,}6\ ;\ 6{,}3\ ;\ -5{,}9\ ;\ 0{,}7\ ;\ -0{,}1\ ;\ 4{,}4\ ;\ -11{,}7\ ;\ 10\ ;\ 2{,}3\ ;\ -3{,}7\ ;\ 5{,}8\ ;\ -1{,}4\ ;$
   $1{,}7\ ;\ -0{,}7$

   Solution:

   a) Below is the box-and-whisker diagram of the data as well as dots representing the data themselves. Note that learners do not neeed to draw the dots, but this helps us to see that there are two outliers on the left.

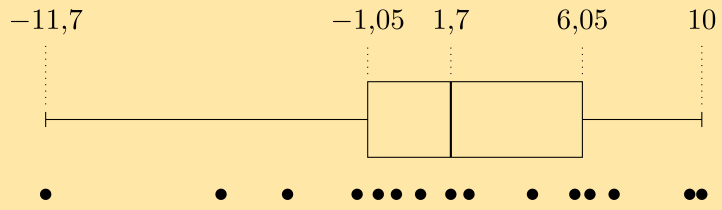
   

   b)

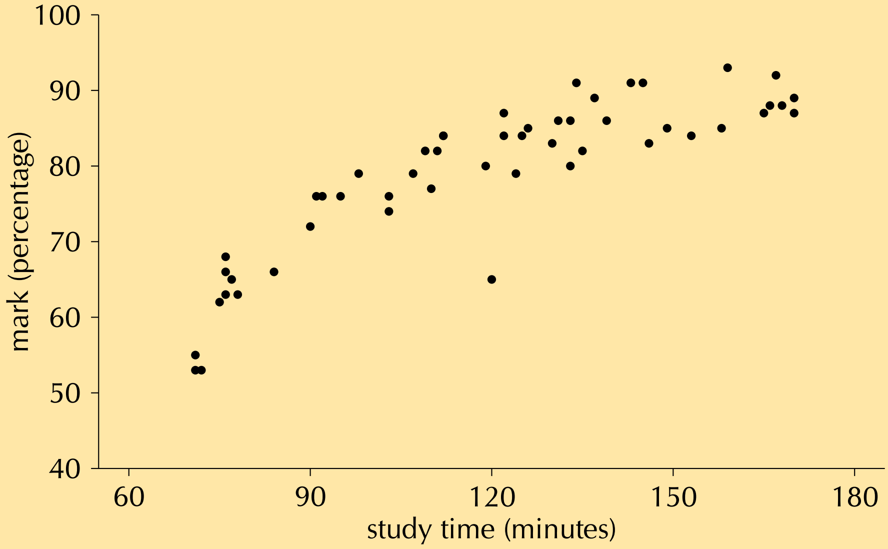
   There are no outliers.

   c)



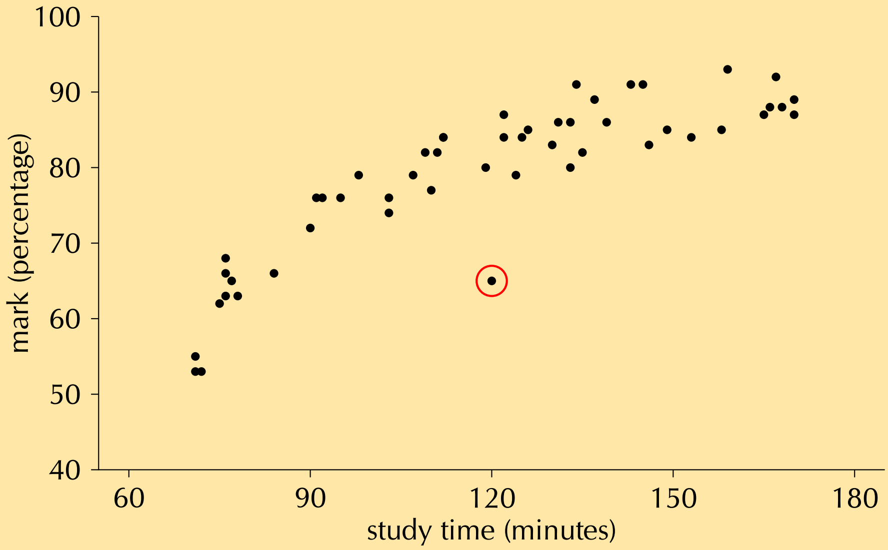
There is one outlier on the left.

2. A class's results for a test were recorded along with the amount of time spent studying for it. The results are given below. Identify any outliers in the data.



Solution:

There is one outlier, marked in red below.



Exercise 11 – 7: End of chapter exercises

1. Draw a histogram, frequency polygon and ogive of the following data set. To count the data, use intervals with a width of 1, starting from 0.

   $0{,}4$; $3{,}1$; $1{,}1$; $2{,}8$; $1{,}5$; $1{,}3$; $2{,}8$; $3{,}1$; $1{,}8$; $1{,}3$;

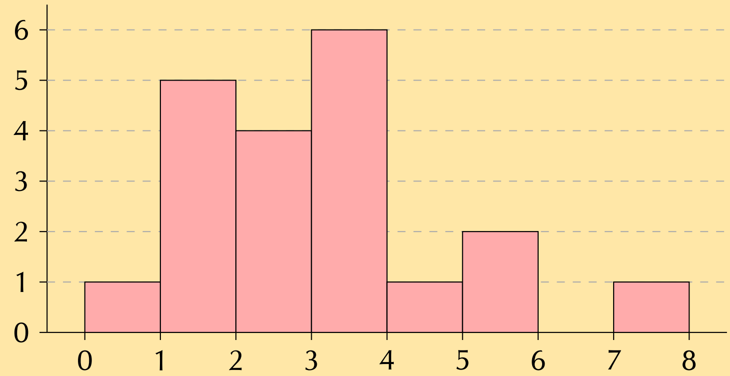
   $2{,}6$; $3{,}7$; $3{,}3$; $5{,}7$; $3{,}7$; $7{,}4$; $4{,}6$; $2{,}4$; $3{,}5$; $5{,}3$

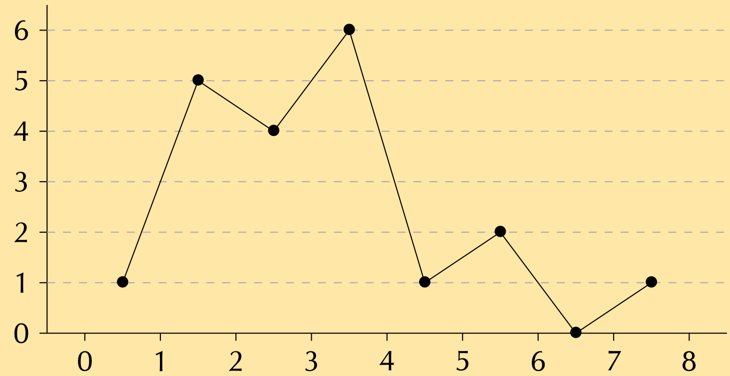
   Solution:

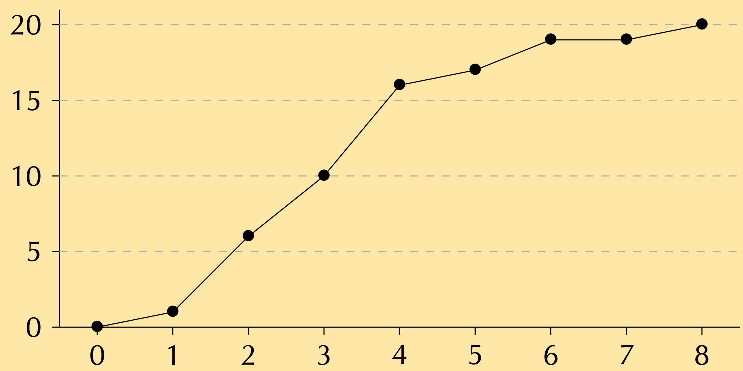
   We first organise the data into a table using an interval width of 1, showing the count in each interval as well as the cumulative count across intervals.

Interval	$[0;1)$	$[1;2)$	$[2;3)$	$[3;4)$	$[4;5)$	$[5;6)$	$[6;7)$	$[7;8)$
Count	1	5	4	6	1	2	0	1
Cumulative	1	6	10	16	17	19	19	20

   From the table above we can draw the histogram, frequency polygon and ogive.







2. Draw a box and whisker diagram of the following data set and explain whether it is symmetric, skewed right or skewed left.

   $-4{,}1$; $-1{,}1$; -1 ; $-1{,}2$; $-1{,}5$; $-3{,}2$; -4 ; $-1{,}9$; -4 ;

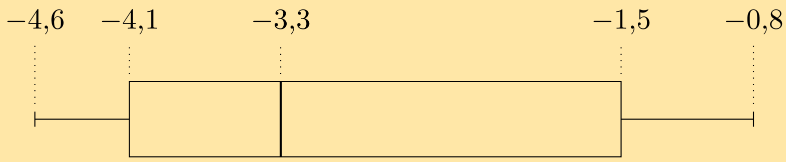
   $-0{,}8$; $-3{,}3$; $-4{,}5$; $-2{,}5$; $-4{,}4$; $-4{,}6$; $-4{,}4$; $-3{,}3$

   Solution:

   The statistics of the data set are

   - minimum: $-4{,}6$;
   - first quartile: $-4{,}1$;
   - median: $-3{,}3$;
   - third quartile: $-1{,}5$;
   - maximum: $-0{,}8$.

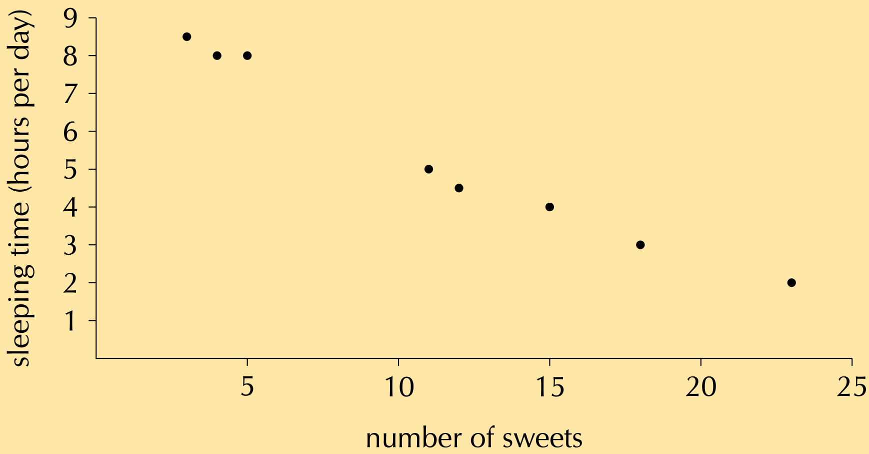
   From this we can draw the box-and-whisker plot as follows.



   Since the median is closer to the first quartile than the third quartile, the data set is skewed right.

3. Eight children's sweet consumption and sleeping habits were recorded. The data are given in the following table and scatter plot.

Number of sweets per week	15	12	5	3	18	23	11	4
Average sleeping time (hours per day)	4	4,5	8	8,5	3	2	5	8



   a) What is the mean and standard deviation of the number of sweets eaten per day?

   b) What is the mean and standard deviation of the number of hours slept per day?

   c) Make a list of all the outliers in the data set.

Solution:

a) Mean $= 11\frac{3}{8}$. Standard deviation $= 6{,}69$.

b) Mean $= 5\frac{3}{8}$. Standard deviation $= 2{,}33$.

c) There are no outliers.

4. The monthly incomes of eight teachers are as follows:

R 10 050; R 14 300; R 9800; R 15 000; R 12 140; R 13 800; R 11 990; R 12 900.

a) What is the mean and standard deviation of their incomes?

b) How many of the salaries are less than one standard deviation away from the mean?

c) If each teacher gets a bonus of R 500 added to their pay what is the new mean and standard deviation?

d) If each teacher gets a bonus of 10% on their salary what is the new mean and standard deviation?

e) Determine for both of the above, how many salaries are less than one standard deviation away from the mean.

f) Using the above information work out which bonus is more beneficial financially for the teachers.

Solution:

a) Mean = R 12 497,50. Standard deviation = R 1768,55.

b) All salaries within the range $(10\ 728{,}95\ ;\ 14\ 266{,}05)$ are less than one standard deviation away from the mean. There are 4 salaries inside this range.

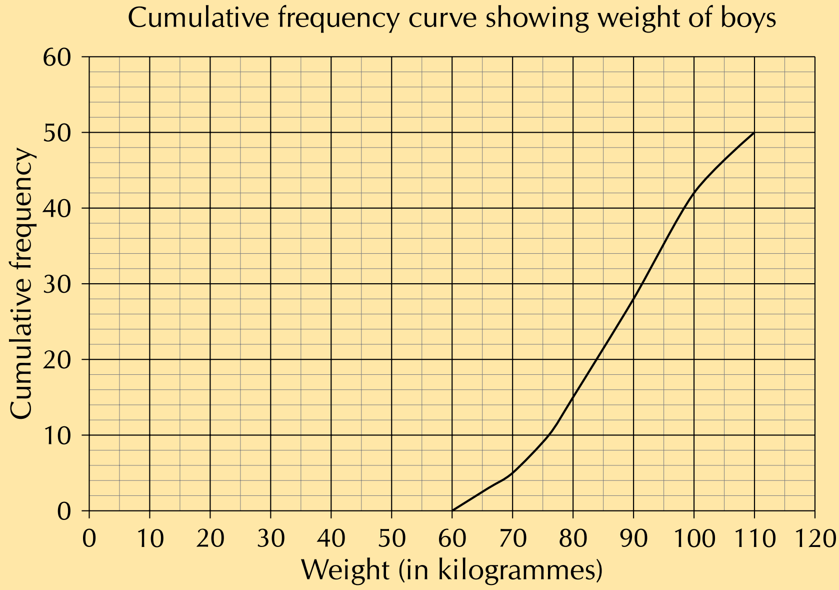
c) Since the increase in each salary is the same absolute amount, the mean simply increases by the bonus. The standard deviation does not change since every value is increased by exactly the same amount. Mean = R 12 997,50. Standard deviation = R 1768,55.

d) With a relative increase, the mean and standard deviation are both multiplied by the same factor. With an increase of 10% the factor is $1{,}1$. Mean = R 13 747,25. Standard deviation = R 1945,41.

e) Adding a constant amount or multiplying by a constant factor (that is, applying a linear transformation) does not change the number of values that lie within one standard deviation from the mean. Therefore the answer is still 4.

f) Since the mean is greater in the second case it means that, on average, the teachers are getting better salaries when the increase is 10%.

5. The weights of a random sample of boys in Grade 11 were recorded. The cumulative frequency graph (ogive) below represents the recorded weights.



a) How many of the boys weighed between 90 and 100 kilogrammes?

b) Estimate the median weight of the boys.

c) If there were 250 boys in Grade 11, estimate how many of them would weigh less than 80 kilogrammes?

Solution:

a) $42 - 28 = 14$

b) There are 50 boys in total, so the median weight is that of the 25^{th} boy. The weight corresponding to a cumulative frequency of 25 is approximately 88 kg.

Note: Accept a range from 86 to 89 kg.

c) 15 boys in the sample have a weight of less than 80 kg. One would expect $\frac{15}{50} \times 250 = 75$ boys in the grade to have a weight of less than 80 kg.

6. Three sets of 12 learners each had their test scores recorded. The test was out of 50. Use the given data to answer the following questions.

Set A	Set B	Set C
25	32	43
47	34	47
15	35	16
17	32	43
16	25	38
26	16	44
24	38	42
27	47	50
22	43	50
24	29	44
12	18	43
31	25	42

a) For each of the sets calculate the mean and the five number summary.

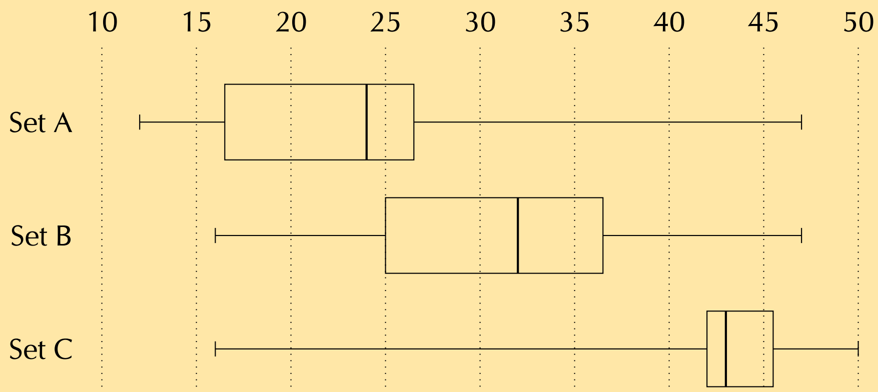
b) Make box and whisker plots of the three data sets on the same set of axes.

c) State, with reasons, whether each of the three data sets are symmetric or skewed (either right or left).

Solution:

a) A. Mean $= 23{,}83$. Five number summary $= [\,12\,;\,16{,}5\,;\,24\,;\,26{,}5\,;\,47\,]$.
B. Mean $= 31{,}17$. Five number summary $= [\,16\,;\,25\,;\,32\,;\,36{,}5\,;\,47\,]$.
C. Mean $= 41{,}83$. Five number summary $= [\,16\,;\,42\,;\,43\,;\,45{,}5\,;\,50\,]$.

b)



c) Set A: skewed left. Set B: slightly skewed left. Set C: skewed right.