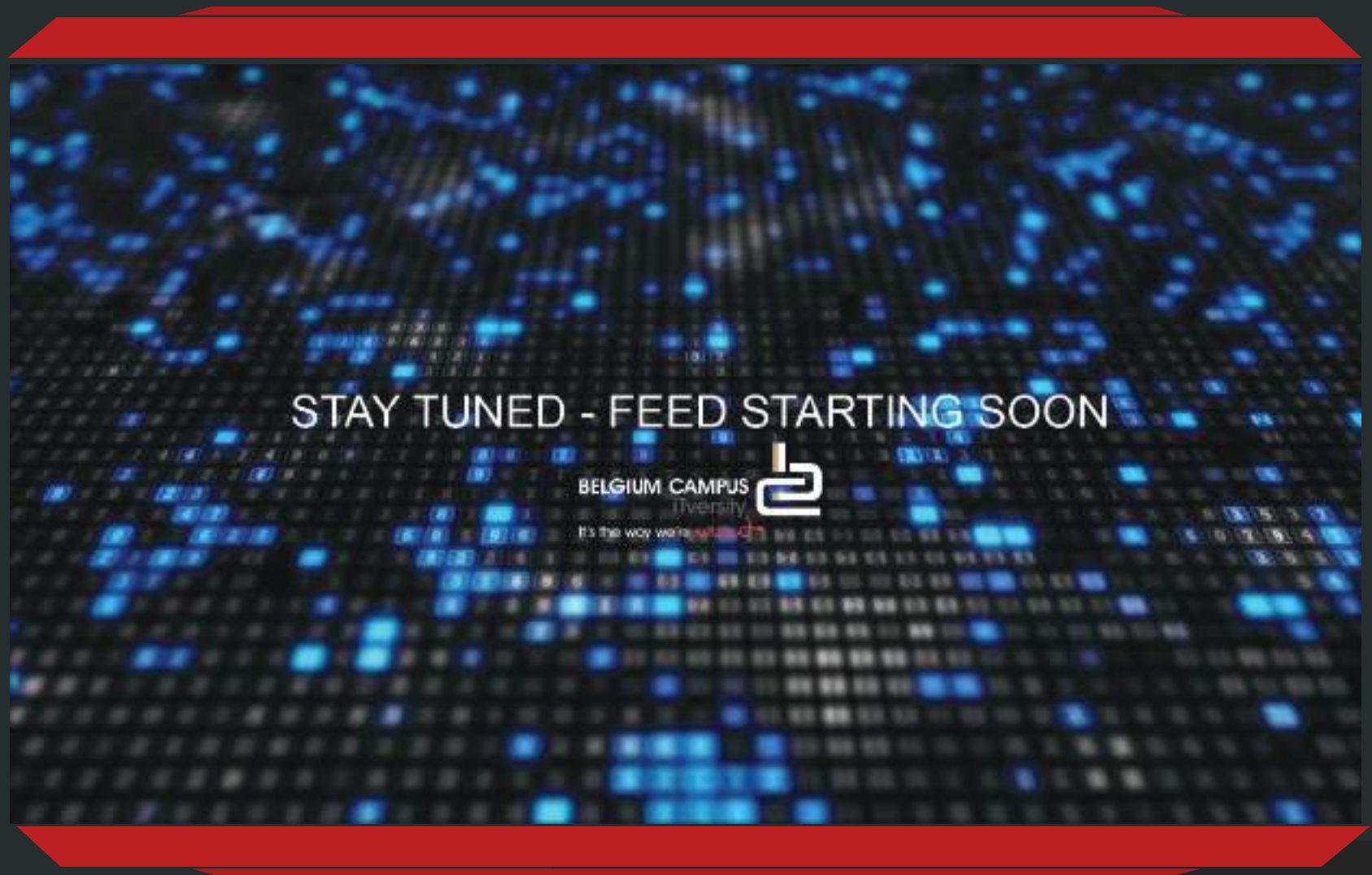

STAY TUNED - FEED STARTING SOON
BELGIUM CAMPUS
iTversity

BELGIUM CAMPUS
iTversity
It's the way we're wired

MATHEMATICS - STATISTICS

E. Cheteni

LESSON OBJECTIVES

- Skewness of data
- Variance
- Curve fitting
- Histograms
- Ogives

SKEWNESS OF DATA

- Symmetrical is when then the median=mean or approximately equal
- Skewed to the right is when the longer tail is to the right median is less than mean
- Skewed to the left is when the longer tail is to the left median is greater than the mean
- To determine skewness of data different diagrams can be used e.g. normal curves, frequency polygons, histograms and box and whisker.

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

a) Mean:

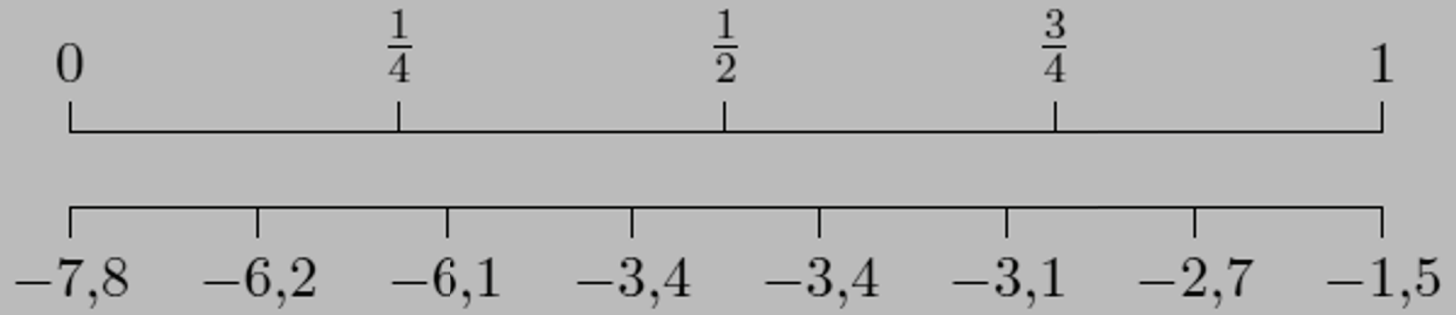
$$\overline{x} = \frac{(-3{,}4) + (-3{,}1) + (-6{,}1) + (-1{,}5) + (-7{,}8) + (-3{,}4) + (-2{,}7) + (-6{,}2)}{8}$$

$$\approx -4{,}3$$

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.



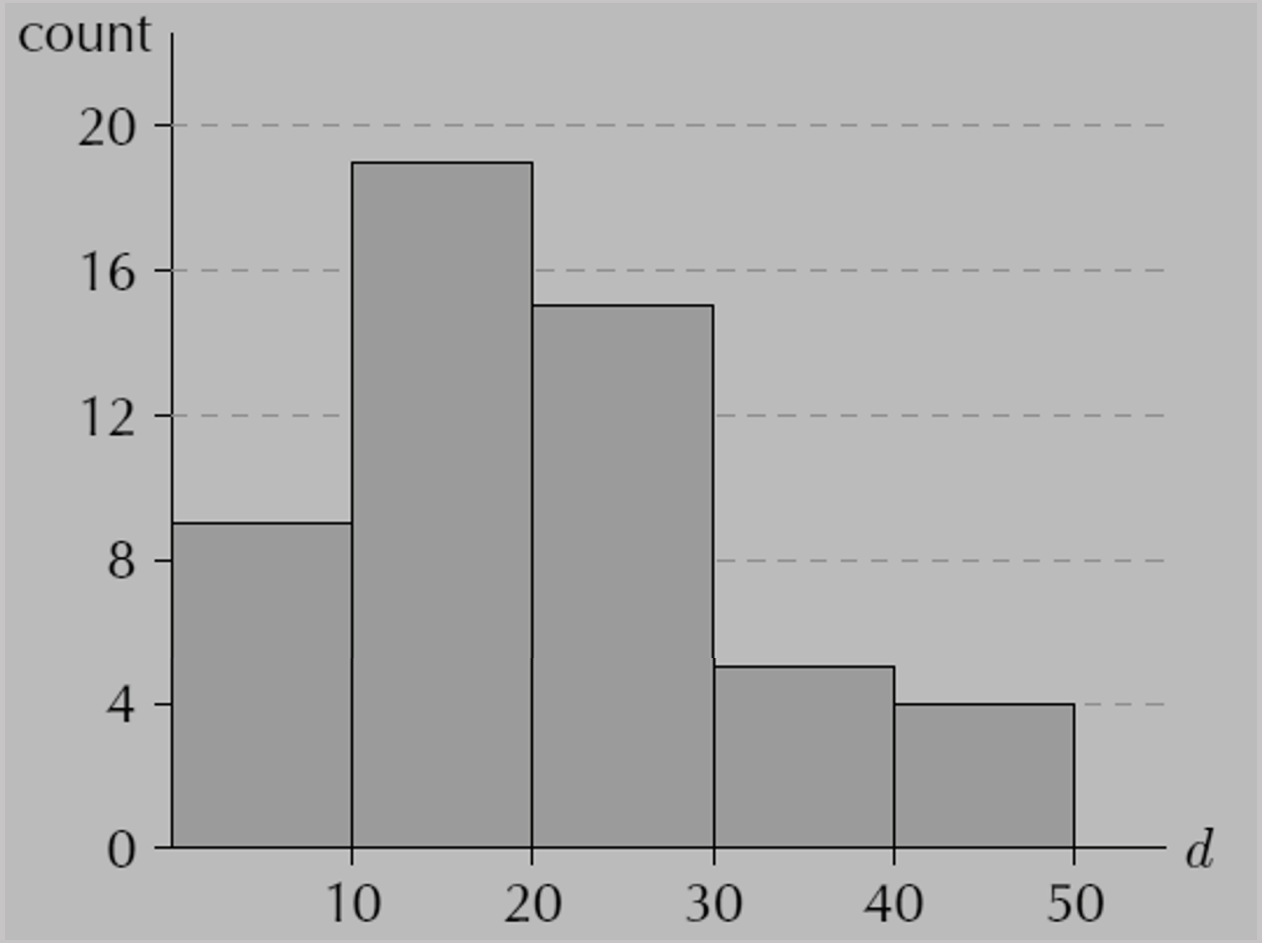
For the first quartile the position is between the second and third values. The second value is $-6{,}2$ and the third value is $-6{,}1$, which means that the first quartile is $\frac{-6{,}2-6{,}1}{2} = -6{,}15$.

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 $-3{,}4$.

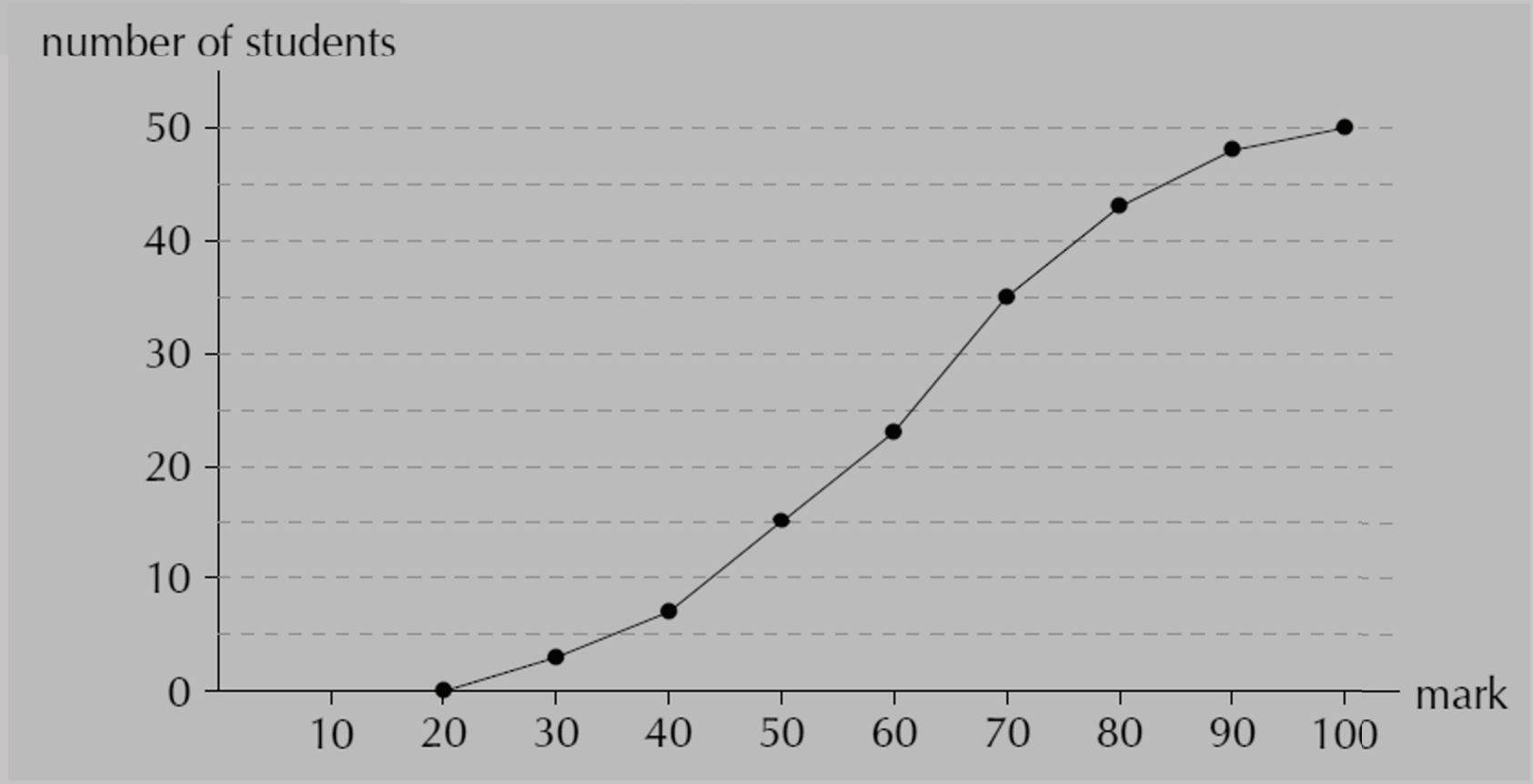
For the third quartile the position is between the sixth and seventh values. Therefore the third quartile is $-2{,}9$.

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

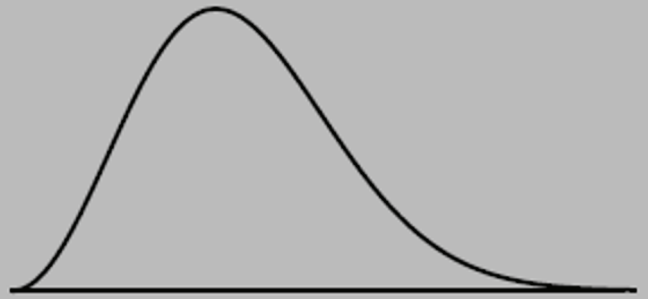

count
20
16
12
8
4
0
10
20
30
40
50
d

Ogive


number of students
50
40
30
20
10
0
10
20
30
40
50
60
70
80
90
100
mark

Examples

a) A data set with this distribution:

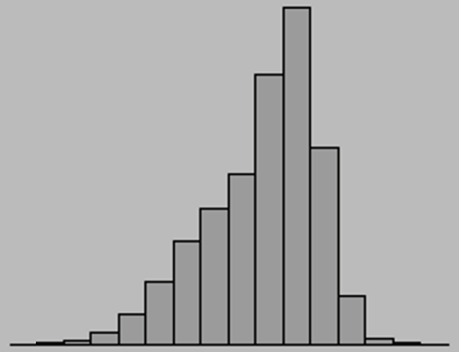


Solution: skewed right

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

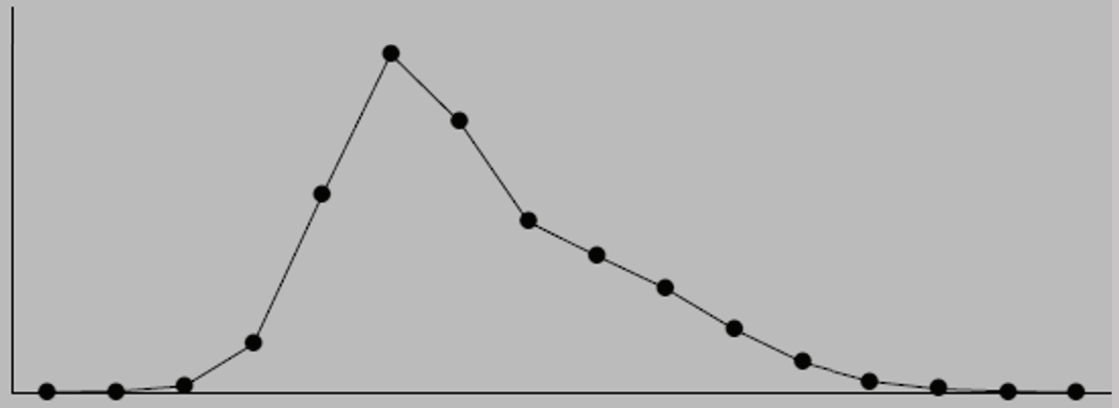
Solution: symmetric

c) A data set with this histogram:



Solution: skewed left

d) A data set with this frequency polygon:



Variance examples

$\{9; 5; 1; 3; 3; 5; 7; 4; 10; 8\}$

Solution:

The formula for the mean is

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

$$\therefore \bar{x} = \frac{55}{10}$$

$$= 5{,}5$$

The formula for the variance is

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

We first subtract the mean from each data point and then square the result.

x_i	9	5	1	3	3	5	7	4	10	8
$x_i - \overline{x}$	3,5	−0,5	−4,5	−2,5	−2,5	−0,5	1,5	−1,5	4,5	2,5
$(x_i - \overline{x})^2$	12,25	0,25	20,25	6,25	6,25	0,25	2,25	2,25	20,25	6,25

Exercise

The variance is the sum of the last row in this table divided by 10, so $\sigma^2 = \frac{76{,}5}{10} = 7{,}65$. The standard deviation is the square root of the variance, therefore $\sigma = \sqrt{7{,}65} = \pm 2{,}77$.

The interval containing all values that are one standard deviation from the mean is $[5{,}5 - 2{,}77; 5{,}5 + 2{,}77] = [2{,}73; 8{,}27]$. We are asked how many values are **within** than one standard deviation from the mean, meaning **inside** the interval. There are 7 values from the data set within the interval, which is $\frac{7}{10} \times 100 = 70\%$ of the data points.

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 ;

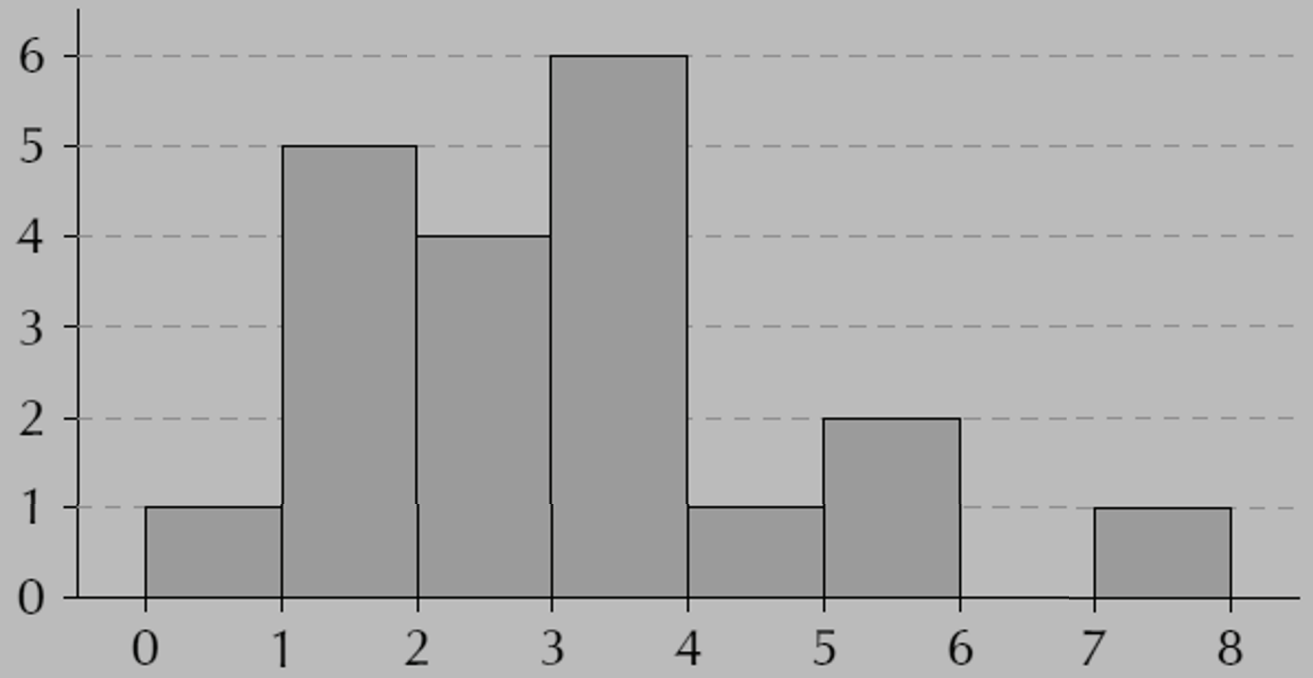
   2,6 ; 3,7 ; 3,3 ; 5,7 ; 3,7 ; 7,4 ; 4,6 ; 2,4 ; 3,5 ; 5,3

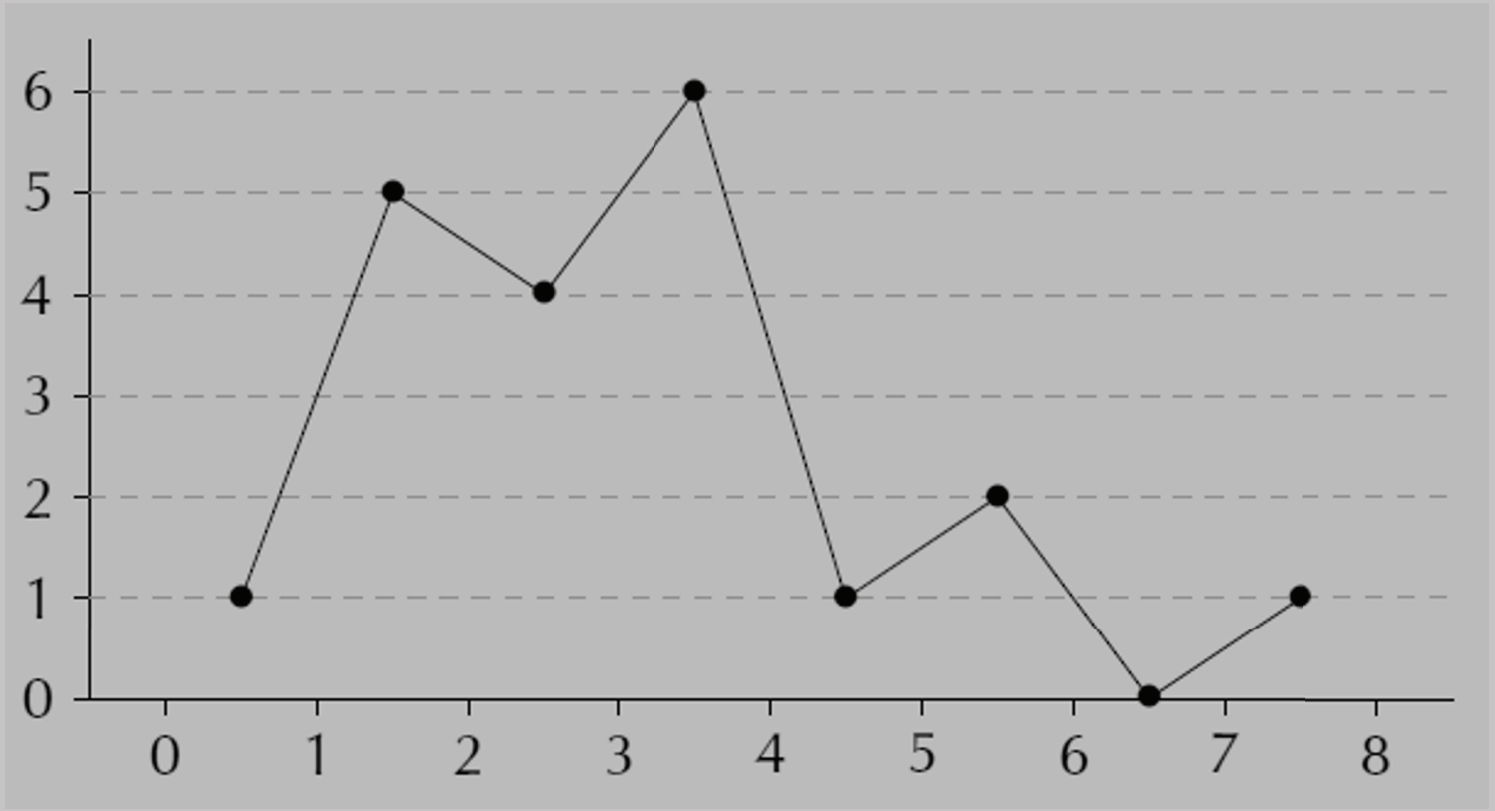
   Solution:

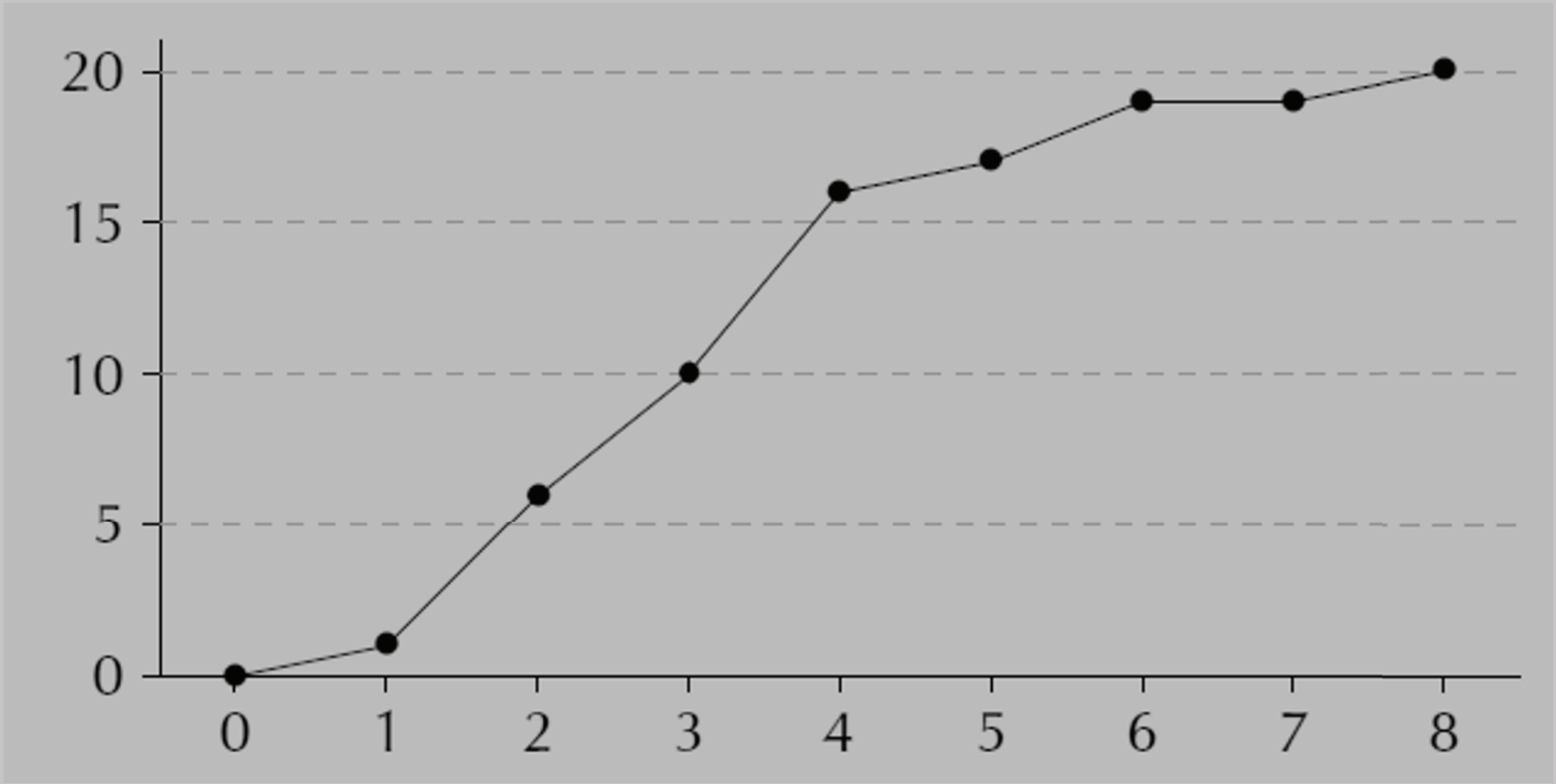
   We first organise the data into a table using an interval width of 1, showing the count in each interval as well as the cumulative count across intervals.

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.




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


20
15
10
5
0
0
1
2
3
4
5
6
7
8

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

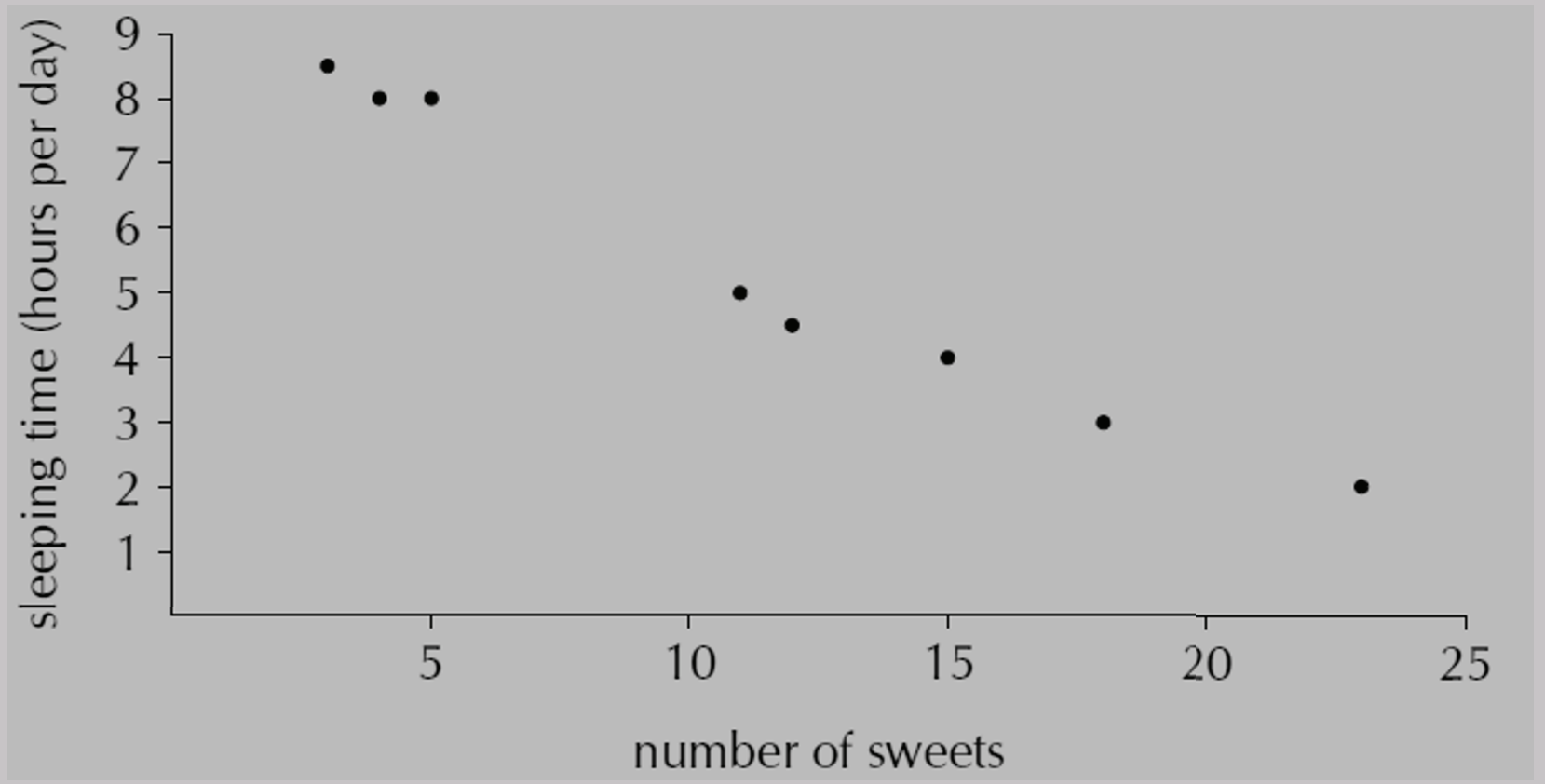
- minimum: $-4{,}6$;
- first quartile: $-4{,}1$;
- median: $-3{,}3$;
- third quartile: $-1{,}5$;
- maximum: $-0{,}8$.

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


sleeping time (hours per day)
9
8
7
6
5
4
3
2
1
5
10
15
20
25
number of sweets

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%.

Questions Grade 11

The table below shows the number of cans of food collected by 9 classes during a charity drive.

5	8	15	20	25	27	31	36	75

1.1 Calculate the range of the data.

1.2 Calculate the standard deviation of the data.

1.3 Determine the median of the data.

1.4 Determine the interquartile range of the data.

1.5 Use the number line provided in the ANSWER BOOK to draw a box and whisker diagram for the data above.

1.6 Describe the skewness of the data.

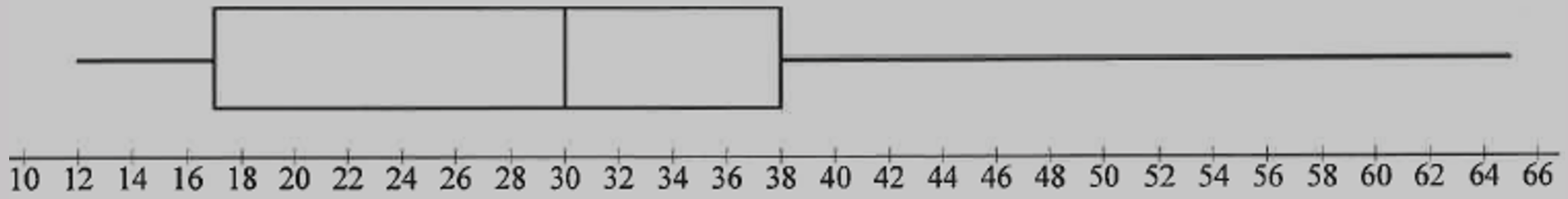
1.7 Identify outliers, if any exist, for the above data.

The table below shows the time (in minutes) that 200 learners spent on their cellphones during a school day.

TIME SPENT (IN MINUTES)	FREQUENCY
$95 < x \leq 105$	15
$105 < x \leq 115$	27
$115 < x \leq 125$	43
$125 < x \leq 135$	52
$135 < x \leq 145$	28
$145 < x \leq 155$	21
$155 < x \leq 165$	10
$165 < x \leq 175$	4

2.1	Complete the cumulative frequency column in the table provided in the ANSWER BOOK.	(2)
2.2	Draw a cumulative frequency graph (ogive) of the data on the grid provided.	(3)
2.3	Use the cumulative frequency graph to determine the value of the lower quartile.	(2)
2.4	Determine, from the cumulative frequency graph, the number of learners who used their cellphones for more than 140 minutes.	(2)

1.1 Mr Brown conducted a survey on the amount of airtime (in rands) EACH student had on his or her cellphone. He summarised the data in the box and whisker diagram below.



1.1.1 Write down the five-number summary of the data. (2)

1.1.2 Determine the interquartile range. (1)

1.1.3 Comment on the skewness of the data. (1)

1.2 A group of 13 students indicated how long it took (in hours) before their cellphone batteries required recharging. The information is given in the table below.

5	8	10	17	20	29	32	48	50	50	63	y	107

1.2.1 Calculate the value of y if the mean for this data set is 41. (2)

1.2.2 If $y = 94$, calculate the standard deviation of the data. (1)

1.2.3 The mean time before another group of 6 students needed to recharge the batteries of their cellphones was 18 hours. Combine these groups and calculate the overall mean time needed for these two groups to recharge the batteries of their cellphones. (3)

A student conducted a survey among his friends and relatives to determine the relationship between the age of a person and the number of marketing phone calls he or she received within one month. The information is given in the table below.

2.1	Complete the frequency and cumulative frequency columns in the table given in the ANSWER BOOK.	(4)
2.2	How many people participated in this survey?	(1)
2.3	Write down the modal class.	(1)
2.4	Draw an ogive (cumulative frequency graph) to represent the data on the grid given in the ANSWER BOOK.	(3)
2.5	Determine the percentage of marketing calls received by people older than 54 years.	(3)

AGE OF PERSON IN SURVEY	FREQUENCY	CUMULATIVE FREQUENCY
$20 < x \leq 30$	7	7
$30 < x \leq 40$		27
$40 < x \leq 50$	25	
$50 < x \leq 60$		64
$60 < x \leq 70$		72
$70 < x \leq 80$	4	
$80 < x \leq 90$		80

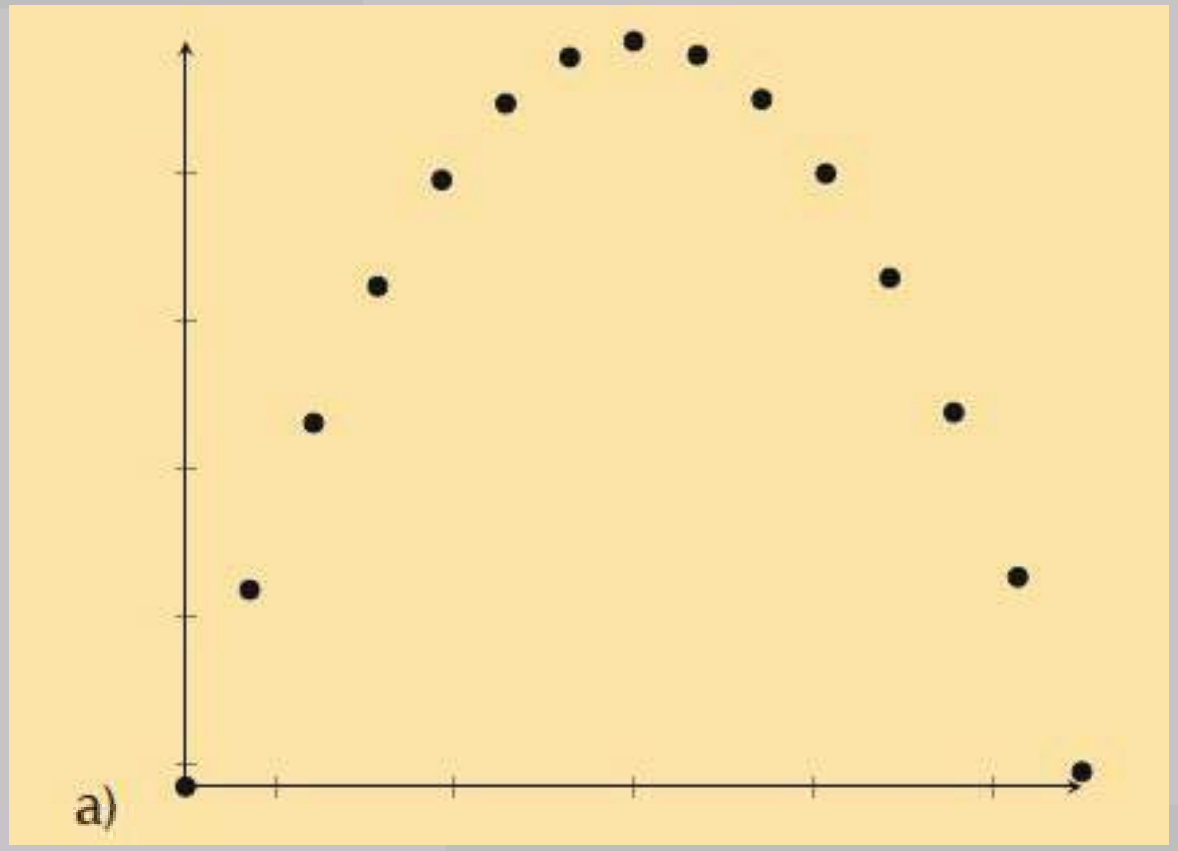
GRADE 12

- Curve Fitting

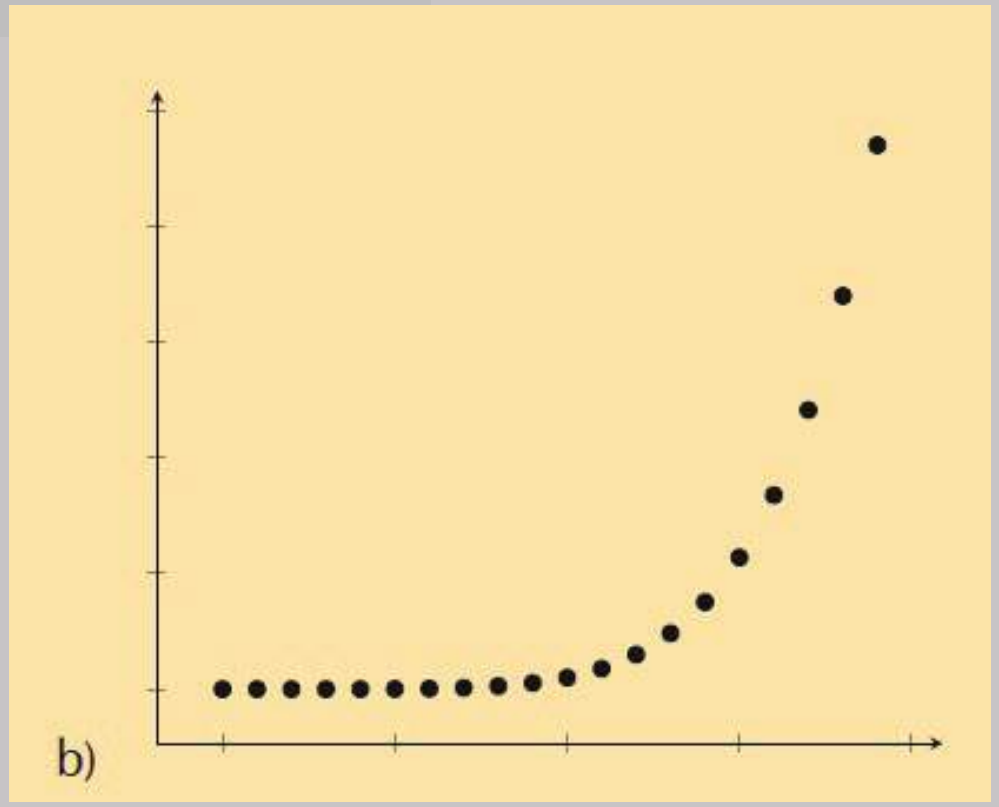
Curve fitting

Identifying the function (linear, exponential or quadratic) which would best fit the data given. This can be achieved through having scatter plots then analyse them.

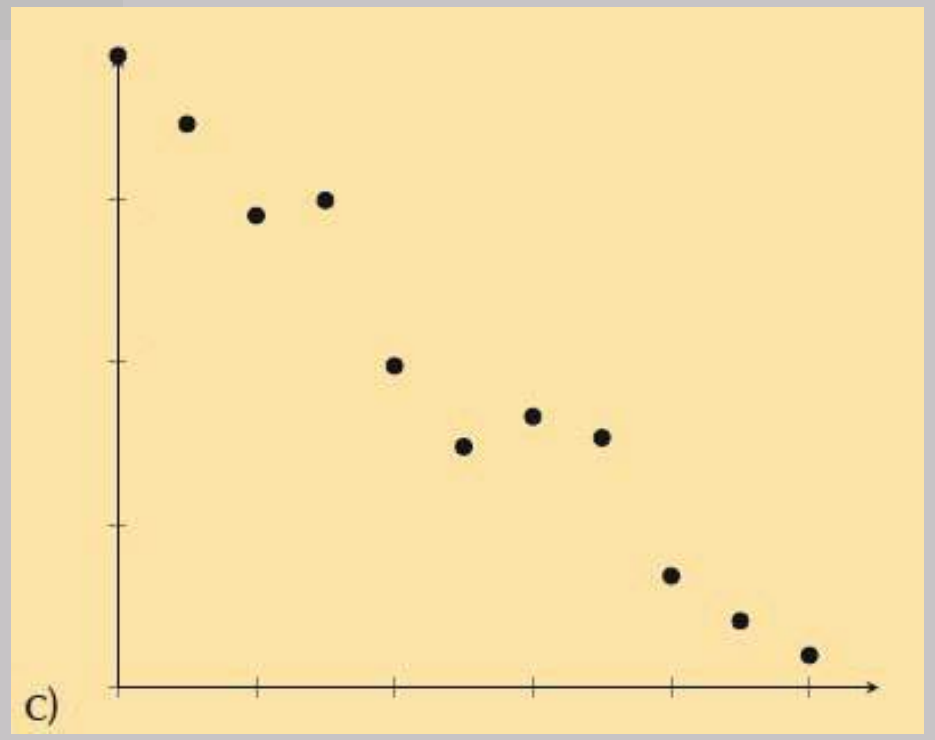
Quadratic



Exponential



Linear

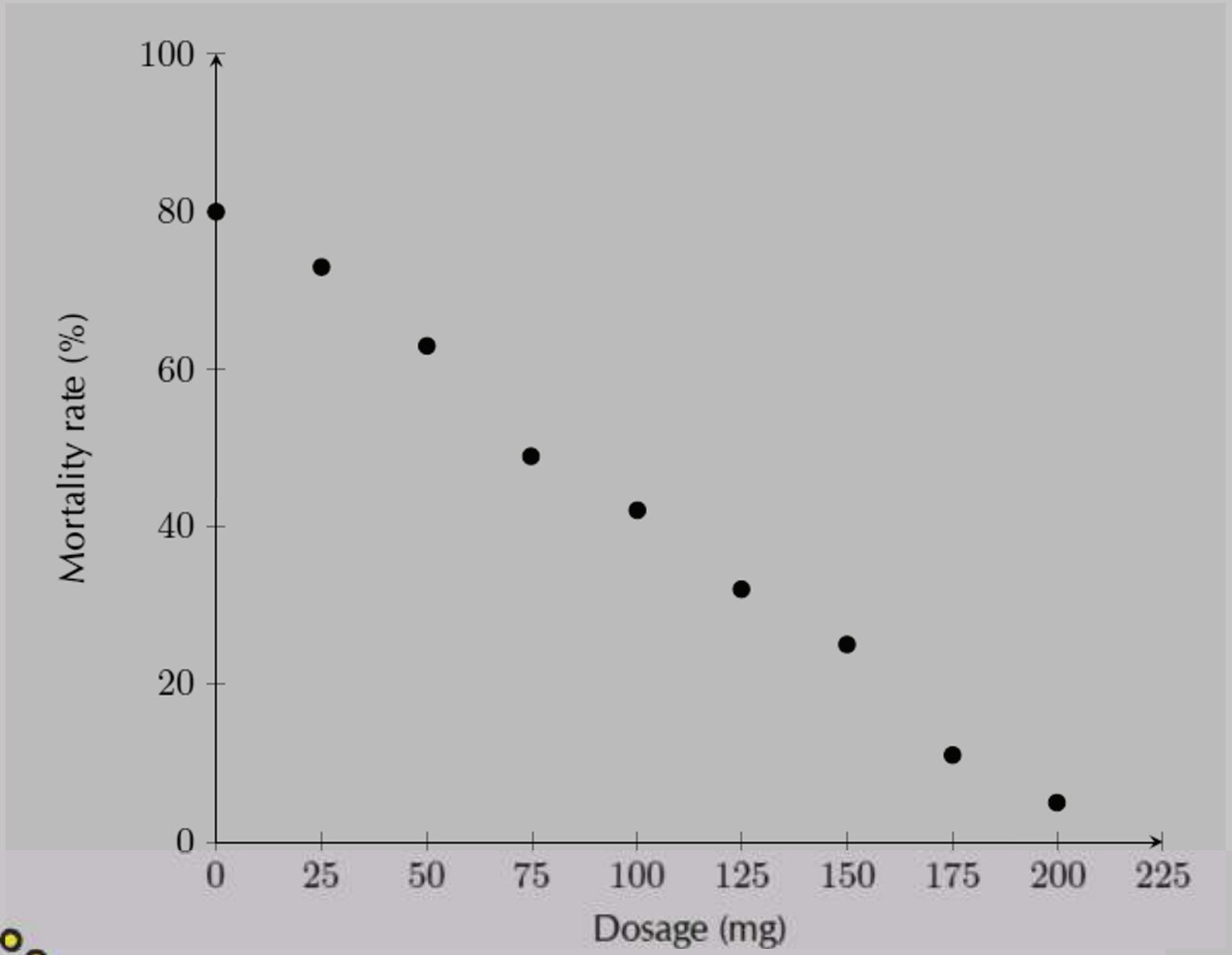


Least squares regression analysis

- Dr Dandara is a scientist trying to find a cure for a disease which has an 80% mortality rate, i.e. 80% of people who get the disease will die. He knows of a plant which is used in traditional medicine to treat the disease. He extracts the active ingredient from the plant and tests different dosages (measured in milligrams) on different groups of patients. Examine his data below and complete the questions that follow.

Dosage (mg)	0	25	50	75	100	125	150	175	200
Mortality rate (%)	80	73	63	49	42	32	25	11	5

a) Draw a scatter plot of the data


100
80
60
40
20
0
Mortality rate (%)
0
25
50
75
100
125
150
175
200
225
Dosage (mg)

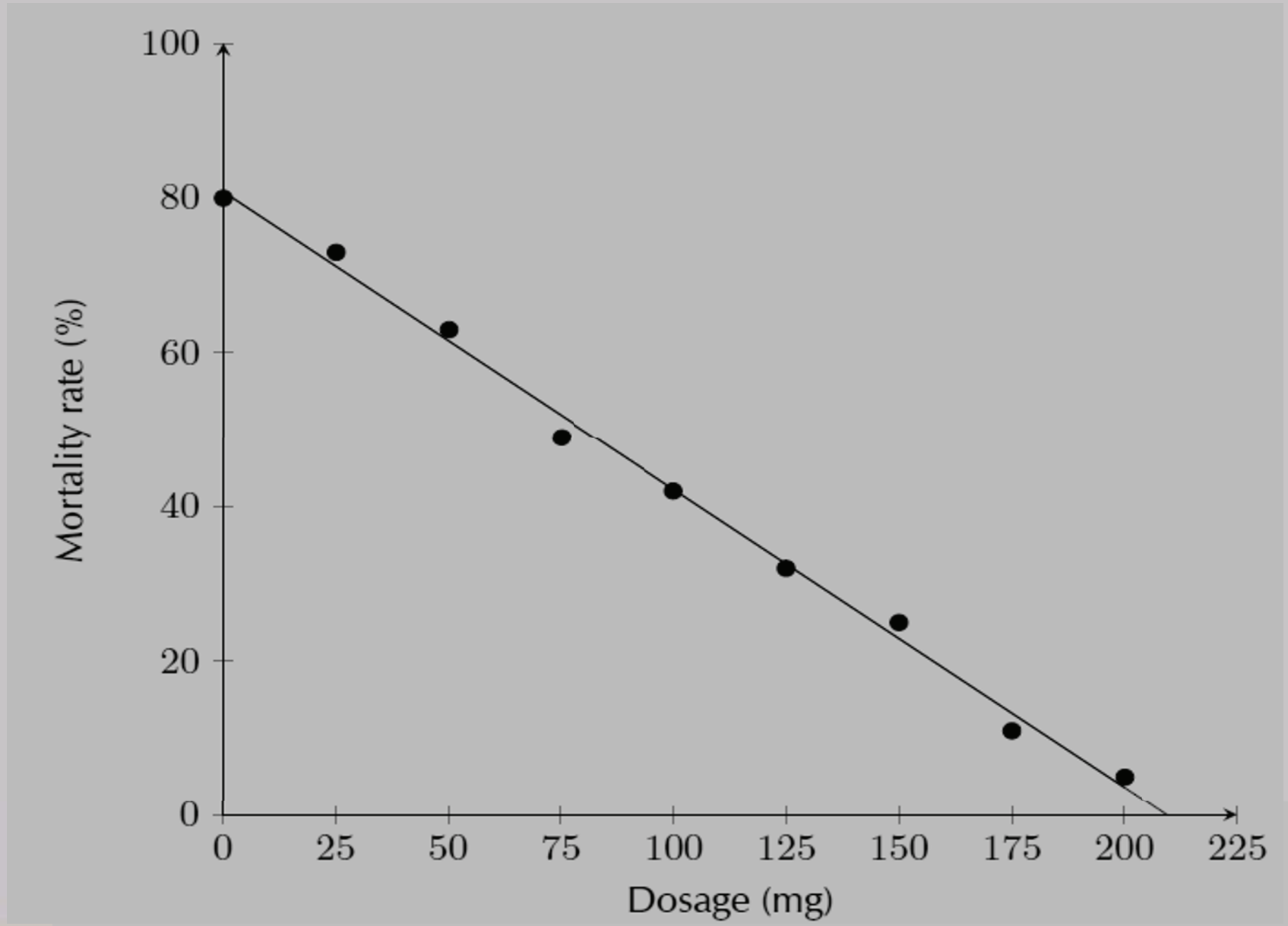
b) Which function would best fit the data? Describe the fit in terms of strength and direction.

Solution:

The data show a strong, negative linear relationship.

c) Draw a line of best fit through the data and determine the equation of your line.

Solution:


100
80
60
40
20
0
Mortality rate (%)
0
25
50
75
100
125
150
175
200
225
Dosage (mg)

The y-intercept is approximately 80. The x-intercept is approximately 210. Therefore, $m = \frac{\Delta y}{\Delta x} = \frac{80-0}{0-210} = -0{,}38$

The equation for the line of best fit: $y = -0{,}38x + 80$

d) Use your equation to estimate the dosage required for a 0% mortality rate.

Solution:

$$0 = -0{,}38x + 80$$

$$\therefore x = \frac{-80}{-0{,}38} = 210{,}53 \text{ mg}$$

e) Dr Dandara decided to administer the estimated dosage required for a 0% mortality rate to a group of infected patients. However, he still found a mortality rate of 5%. Name the statistical technique Dr Dandara used to estimate a mortality rate of 0% and explain why his equation did not accurately predict his experimental results.

Solution:

Dr Dandara used **extrapolation** to calculate the dosage where the mortality rate $= 0\%$. Extrapolation can result in incorrect estimates if the trend observed within the available data range does not continue outside of the range. In this case, it appears that at dosages greater than 200 mg, the equation of the line of best fit no longer fits the data, therefore extrapolation produced a false estimate.

Linear regression

Determine the equation of the least-squares regression line using a table for the data sets below. Round **a** and **b** to two decimal places.

a)

x	10	4	9	11	11	6	8	18	9	13
y	1	0	6	3	9	5	9	8	7	15

Solution

x	y	xy	x^2
10	1	10	100
4	0	0	16
9	6	54	81
11	3	33	121
11	9	99	121
6	5	30	36
8	9	72	64
18	8	144	324
9	7	63	81
13	15	195	169
$\sum = 99$	$\sum = 63$	$\sum = 700$	$\sum = 1113$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 \times 700 - 99 \times 63}{10 \times 1113 - 99^2} = 0{,}574$$

$$a = \bar{y} - b\bar{x} = \frac{63}{10} - \frac{0{,}574 \times 99}{10} = 0{,}616$$

$$\therefore \hat{y} = 0{,}62 + 0{,}57x$$

Example

Determine the equation of the least squares regression line given each set of data values below. Round a and b to two decimal places in your final answer.

a) $n = 10;\ \sum x = 74;\ \sum y = 424;\ \sum xy = 4114{,}51;\ \sum(x^2) = 718{,}86$

Solution

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{10 \times 4114{,}51 - 74 \times 424}{10 \times 718{,}86 - 74^2} = 5{,}704250847$$

$$a = \bar{y} - b\bar{x} = \frac{424}{10} - 5{,}704250847 \times \frac{74}{10} = 0{,}188543732$$

$$\therefore \hat{y} = 0{,}19 + 5{,}70x$$

Correlation coefficient

- Determine the correlation coefficient by hand for the following data sets and comment on the strength and direction of the correlation. Round your answers to two decimal places.

a)

x	5	8	13	10	14	15	17	12	18	13
y	5	8	3	8	7	5	3	-1	4	-1

x	y	xy	x^2	$x - \bar{x}^2$	$y - \bar{y}^2$
5	5	25	25	56,25	0,81
8	8	64	64	20,25	15,21
13	3	39	169	0,25	1,21
10	8	80	100	6,25	15,21
14	7	98	196	2,25	8,41
15	5	75	225	6,25	0,81
17	3	51	289	20,25	1,21
12	−1	−12	144	0,25	26,01
18	4	72	324	30,25	0,01
13	−1	−13	169	0,25	26,01
$\sum =$ 125	$\sum =$ 41	$\sum =$ 479	$\sum =$ 1705	$\sum =$ 142,5	$\sum =$ 94,9

$$r = b\frac{\sigma_x}{\sigma_y}$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 \times 479 - 125 \times 41}{10 \times 1705 - 125^2} = -0{,}235$$

$$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{142{,}5}{10}} = \sqrt{14{,}25} = \pm 3{,}775$$

$$\sigma_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n}} = \sqrt{\frac{94{,}9}{10}} = \sqrt{9{,}49} = \pm 3{,}081$$

$$\therefore r = -0{,}235 \times \frac{3{,}775}{3{,}081}$$

$$= -0{,}29$$

Therefore, the correlation between x and y is negative but weak.

Exercise Grade 12

- Learners who scored a mark below 50% in a Mathematics test were selected to use a computer-based programme as part of an intervention strategy. On completing the programme, these learners wrote a second test to determine the effectiveness of the intervention strategy. The mark (as a percentage) scored by 15 of these learners in both tests is given in the table below.

LEARNER	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15
TEST 1 (%)	10	18	23	24	27	34	34	36	37	39	40	44	45	48	49
TEST 2 (%)	33	21	32	20	58	43	49	48	41	55	50	45	62	68	60

2.1 Determine the equation of the least squares regression line. (3)

2.2 A learner's mark in the first test was 15 out of a maximum of 50 marks.

2.2.1 Write down the learner's mark for this test as a percentage. (1)

2.2.2 Predict the learner's mark for the second test. Give your answer to the nearest integer. (2)

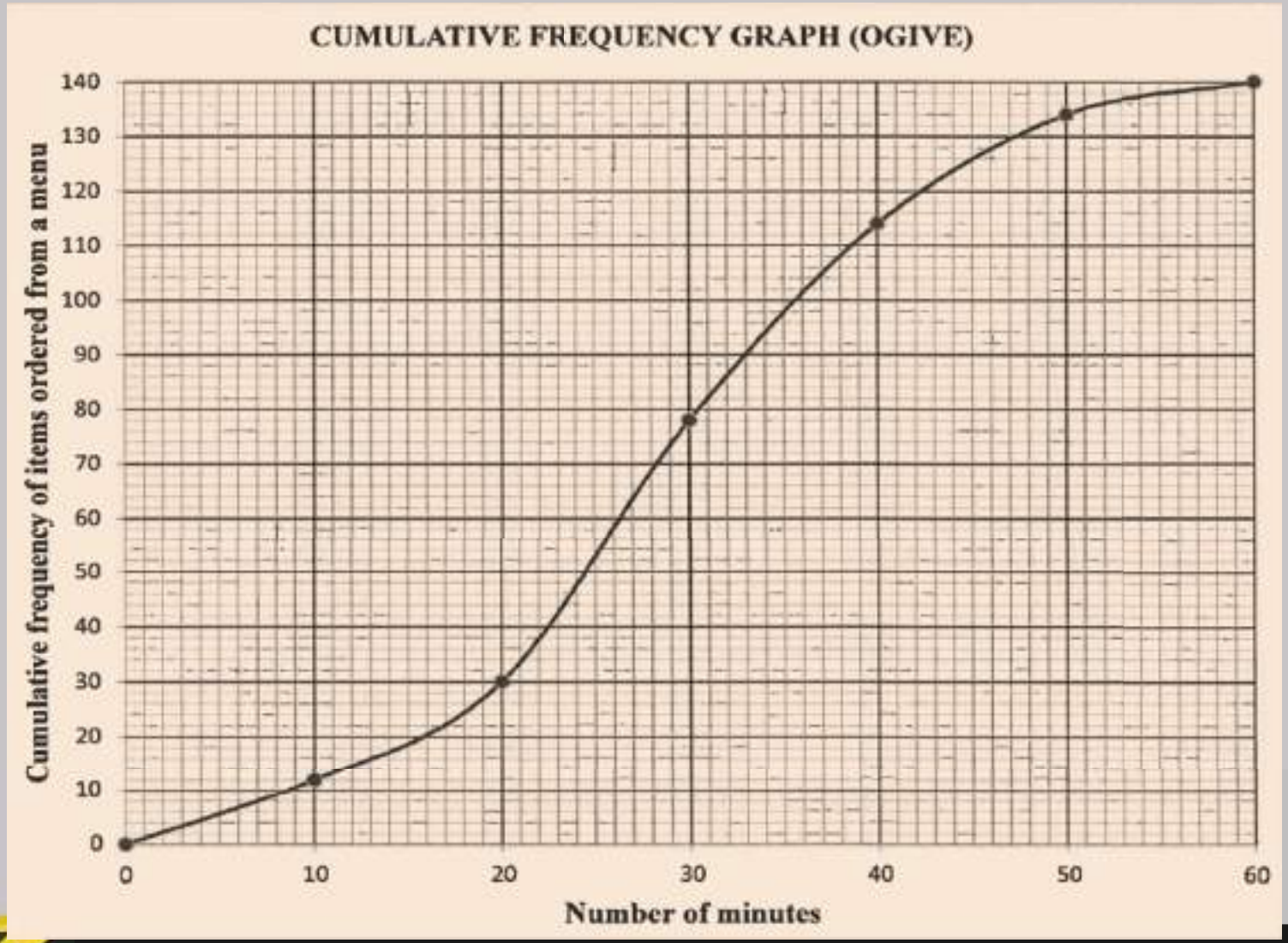
2.3 For the 15 learners above, the mean mark of the second test is 45,67% and the standard deviation is 13,88%. The teacher discovered that he forgot to add the marks of the last question to the total mark of each of these learners. All the learners scored full marks in the last question. When the marks of the last question are added, the new mean mark is 50,67%.

2.3.1 What is the standard deviation after the marks for the last question are added to each learner's total? (2)

2.3.2 What is the total mark of the last question? (2)

[10]

The cumulative frequency graph (ogive) drawn below shows the total number of food items ordered from a menu over a period of 1 hour.



1.1.1	Write down the total number of food items ordered from the menu during this hour.	(1)
1.1.2	Write down the modal class of the data.	(1)
1.1.3	How long did it take to order the first 30 food items?	(1)
1.1.4	How many food items were ordered in the last 15 minutes?	(2)
1.1.5	Determine the 75^{th} percentile for the data.	(2)
1.1.6	Calculate the interquartile range of the data.	(2)

1.2 Reggie works part-time as a waiter at a local restaurant. The amount of money (in rands) he made in tips over a 15-day period is given below.

35	70	75	80	80
90	100	100	105	105
110	110	115	120	125

1.2.1 Calculate:

(a) The mean of the data (2)

(b) The standard deviation of the data (2)

1.2.2 Mary also works part-time as a waitress at the same restaurant. Over the same 15-day period Mary collected the same mean amount in tips as Reggie, but her standard deviation was R14.

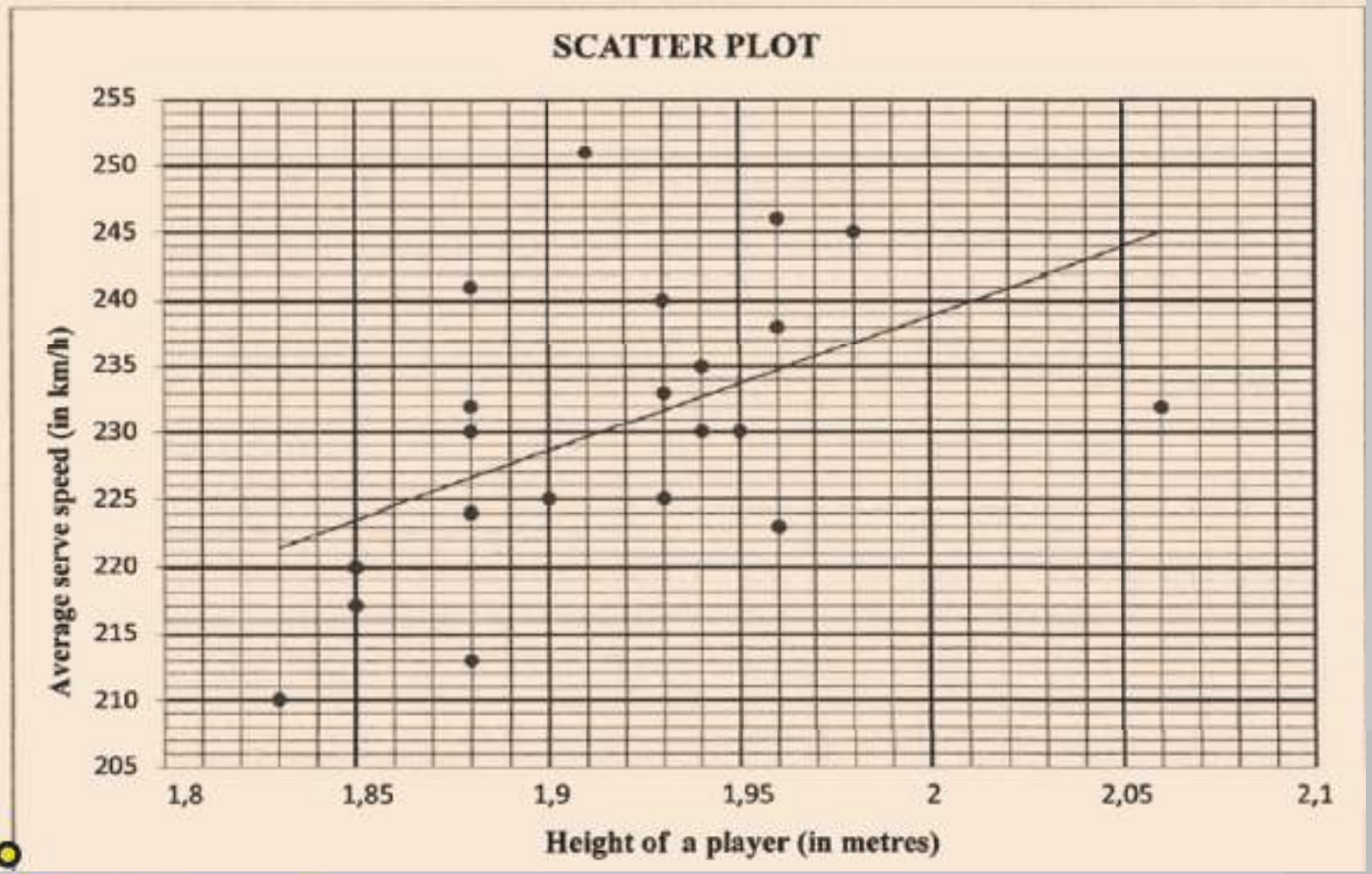
Using the available information, comment on the:

(a) Total amount in tips that they EACH collected over the 15-day period (1)

(b) Variation that EACH of them received in daily tips over this period (1)

[15]

A familiar question among professional tennis players is whether the speed of a tennis serve (in km/h) depends on the height of a player (in metres). The heights of 21 tennis players and the average speed of their serves were recorded during a tournament. The data is represented in the scatter plot below. The least squares regression line is also drawn.



2.1 Write down the fastest average serve speed (in km/h) achieved in this tournament. (1)

2.2 Consider the following correlation coefficients:

A. $r = 0{,}93$ B. $r = -0{,}42$ C. $r = 0{,}52$

2.2.1 Which ONE of the given correlation coefficients best fits the plotted data? (1)

2.2.2 Use the scatter plot and least squares regression line to motivate your answer to QUESTION 2.2.1. (1)

2.3 What does the data suggest about the speed of a tennis serve (in km/h) and the height of a player (in metres)? (1)

2.4 The equation of the regression line is given as $\hat{y} = 27{,}07 + bx$.
Explain why, in this context, the least squares regression line CANNOT intersect the y-axis at $(0\ ;\ 27{,}07)$. (1)
[5]

A survey was conducted among 100 people about the amount that they paid on a monthly basis for their cellphone contracts. The person carrying out the survey calculated the estimated mean to be R309 per month. Unfortunately, he lost some of the data thereafter. The partial results of the survey are shown in the frequency table below:

2.1	How many people paid R200 or less on their monthly cellphone contracts?	(1)
2.2	Use the information above to show that $a = 24$ and $b = 16$.	(5)
2.3	Write down the modal class for the data.	(1)
2.4	On the grid provided in the ANSWER BOOK, draw an ogive (cumulative frequency graph) to represent the data.	(4)
2.5	Determine how many people paid more than R420 per month for their cellphone contracts.	(2)

AMOUNT PAID (IN RANDS)	FREQUENCY
$0 < x \leq 100$	7
$100 < x \leq 200$	12
$200 < x \leq 300$	a
$300 < x \leq 400$	35
$400 < x \leq 500$	b
$500 < x \leq 600$	6