Instructions

Answer **all** questions in the spaces provided.
In all questions where a numerical answer is required, you should only round your answer when instructed to do so.
Unless otherwise indicated, the diagrams in this book are **not** drawn to scale.

Data analysis

Question 1 (4 marks)

The table below displays the *average sleep time*, in hours, for a sample of 19 types of mammals.

Type of mammal	*Average sleep time* **(hours)**
cat	14.5
squirrel	13.8
mouse	13.2
rat	13.2
grey wolf	13.0
arctic fox	12.5
raccoon	12.5
gorilla	12.0
jaguar	10.8
baboon	9.8
red fox	9.8
rabbit	8.4
guinea pig	8.2
grey seal	6.2
cow	3.9
sheep	3.8
donkey	3.1
horse	2.9
roedeer	2.6

Data: T Allison and DV Cicchetti, 'Sleep in Mammals: Ecological and Constitutional Correlates', in *Science*, American Association for the Advancement of Science, vol. 194, no. 4266, pp. 732–734, 12 November 1976; accessed from OzDASL, StatSci.org, <www.statsci.org/data/general/sleep.html>

DO NOT WRITE IN THIS AREA

Question 1 – continued

DO NOT WRITE IN THIS AREA

a. Which of the two variables, *type of mammal* or *average sleep time*, is a nominal variable? 1 mark

b. Determine the mean and standard deviation of the variable *average sleep time* for this sample of mammals.

Write your answers in the boxes provided below.

Round your answers to one decimal place. 1 mark

mean = ☐ hours standard deviation = ☐ hours

c. The average sleep time for a human is eight hours.

What percentage of this sample of mammals has an *average sleep time* that is less than the average sleep time for a human?

Round your answer to one decimal place. 1 mark

d. The sample is increased in size by adding in the average sleep time of the little brown bat. Its average sleep time is 19.9 hours.

By how many hours will the range for *average sleep time* increase when the average sleep time for the little brown bat is added to the sample? 1 mark

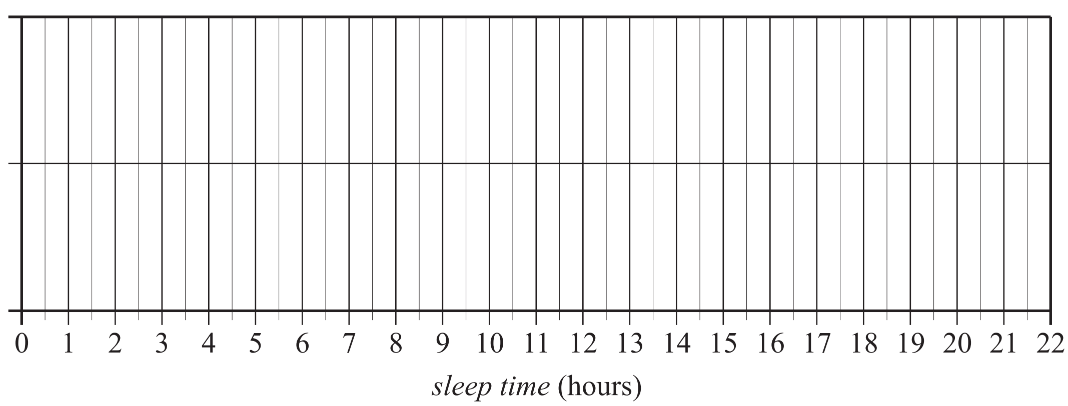
Question 2 (3 marks)

The five-number summary below was determined from the *sleep time*, in hours, of a sample of 59 types of mammals.

Statistic	***Sleep time* (hours)**
minimum	2.5
first quartile	8.0
median	10.5
third quartile	13.5
maximum	20.0

a. Show, with calculations, that a boxplot constructed from this five-number summary will not include outliers. 2 marks

b. Construct the boxplot below. 1 mark



DO NOT WRITE IN THIS AREA

Question 3 (4 marks)

The *life span*, in years, and *gestation period*, in days, for 19 types of mammals are displayed in the table below.

Life span **(years)**	***Gestation period*** **(days)**
3.20	19
4.70	21
7.60	68
9.00	28
9.80	52
13.7	63
14.0	60
16.2	63
17.0	150
18.0	31
20.0	151
22.4	100
27.0	180
28.0	63
30.0	281
39.3	252
40.0	365
41.0	310
46.0	336

a. A least squares line that enables *life span* to be predicted from *gestation period* is fitted to this data.

Name the explanatory variable in the equation of this least squares line. 1 mark

---

b. Determine the equation of the least squares line in terms of the variables *life span* and *gestation period*.

Write your answers in the appropriate boxes provided below.

Round the numbers representing the intercept and slope to three significant figures. 2 marks

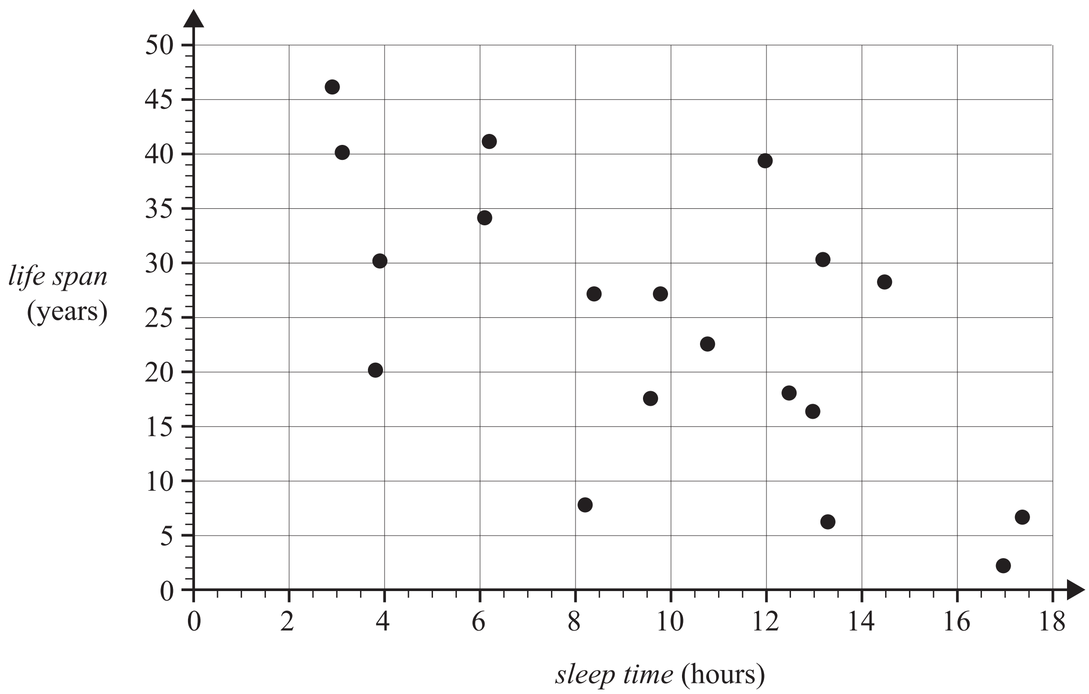
[ ] = [ ] + [ ] × [ ]

c. Write the value of the correlation coefficient rounded to three decimal places. 1 mark

$r =$ [ ]

Question 4 (8 marks)

The scatterplot below plots the variable *life span*, in years, against the variable *sleep time*, in hours, for a sample of 19 types of mammals.



On the assumption that the association between *sleep time* and *life span* is linear, a least squares line is fitted to this data with *sleep time* as the explanatory variable.

The equation of this least squares line is

$$\textit{life span} = 42.1 - 1.90 \times \textit{sleep time}$$

The coefficient of determination is 0.416

a. Draw the graph of the least squares line on the scatterplot above. 1 mark

(*Answer on the scatterplot above.*)

b. Describe the linear association between *life span* and *sleep time* in terms of strength and direction. 2 marks

c. Interpret the slope of the least squares line in terms of *life span* and *sleep time*. 2 marks

DO NOT WRITE IN THIS AREA

Question 4 – continued

DO NOT WRITE IN THIS AREA

d. Interpret the coefficient of determination in terms of *life span* and *sleep time*. 1 mark

e. The life span of the mammal with a sleep time of 12 hours is 39.2 years.

Show that, when the least squares line is used to predict the life span of this mammal, the residual is 19.9 years. 2 marks

Question 5 (5 marks)

A random sample of 12 mammals drawn from a population of 62 types of mammals was categorised according to two variables:

- *likelihood of attack* (low = 1, medium = 2, high = 3)
- *exposure to attack during sleep* (low = 1, medium = 2, high = 3)

The data is shown in the following table.

Likelihood of attack	2	2	1	3	2	3	1	3	1	1	3	3
Exposure to attack during sleep	3	1	1	1	3	3	1	3	1	1	3	3

a. Use this data to complete the two-way frequency table below. 1 mark

Likelihood of attack	***Exposure to attack during sleep***		
	low (=1)	medium (=2)	high (=3)
low (=1)		0	0
medium (=2)		0	
high (=3)		0	

Question 5 – continued

DO NOT WRITE IN THIS AREA

b. The following two-way frequency table was formed from the data generated when the entire population of 62 types of mammals was similarly categorised.

Likelihood of attack	***Exposure to attack during sleep***		
	low (=1)	medium (=2)	high (=3)
low (=1)	31	8	2
medium (=2)	2	0	2
high (=3)	1	1	15

i. How many of these 62 mammals had both a high *likelihood of attack* and a high *exposure to attack during sleep*? 1 mark

ii. Of those mammals that had a medium *likelihood of attack*, what percentage also had a low *exposure to attack during sleep*? 1 mark

iii. Does the information in the table above support the contention that *likelihood of attack* is associated with *exposure to attack during sleep*? Justify your answer by quoting appropriate percentages. It is sufficient to consider only one category of *likelihood of attack* when justifying your answer. 2 marks

Instructions

Answer **all** questions in the spaces provided.
In all questions where a numerical answer is required, you should only round your answer when instructed to do so.
Unless otherwise indicated, the diagrams in this book are **not** drawn to scale.

Data analysis

Question 1 (9 marks)

Data was collected to investigate the use of electronic images to automate the sizing of oysters for sale.

The variables in this study were:

- *ID:* identity number of the oyster
- *weight:* weight of the oyster in grams (g)
- *volume:* volume of the oyster in cubic centimetres (cm^3)
- *image size:* oyster size determined from its electronic image (in megapixels)
- *size:* oyster size when offered for sale: small, medium or large

The data collected for a sample of 15 oysters is displayed in Table 1.

Table 1

ID	*Weight* (g)	*Volume* (cm^3)	*Image size* (megapixels)	*Size*
1	12.9	13.0	5.1	large
2	11.4	11.7	4.8	medium
3	17.4	17.4	6.5	large
4	6.8	7.2	2.9	small
5	9.6	10.1	3.7	medium
6	15.5	15.6	5.7	large
7	9.7	9.9	4.0	small
8	7.0	7.5	2.7	small
9	12.6	12.7	5.5	medium
10	12.5	12.7	5.0	medium
11	10.1	10.5	3.9	medium
12	10.6	10.8	4.1	medium
13	13.0	13.1	5.3	large
14	8.1	8.5	3.5	small
15	14.1	14.2	5.3	large

Data: http://jse.amstat.org/jse_data_archive.htm

a. Write down the number of categorical variables in Table 1. 1 mark

____

Question 1 – continued

DO NOT WRITE IN THIS AREA

b. Determine, in grams:

i. the mean *weight* of all the oysters in this sample 1 mark

mean = ☐

ii. the median *weight* of the large oysters in this sample. 1 mark

median = ☐

c. When a least squares line is used to model the association between oyster *weight* and *volume*, the equation is:

$$volume = 0.780 + 0.953 \times weight$$

i. Name the response variable in this equation. 1 mark

---

ii. Complete the following sentence by filling in the box provided. 1 mark

This equation predicts that, on average, each 10 g increase in the *weight* of an oyster is associated with a ☐ cm^3 increase in its *volume*.

d. A least squares line can also be used to model the association between an oyster's *volume*, in cm^3, and its electronic *image size*, in megapixels. In this model, *image size* is the explanatory variable.

Using data from Table 1, determine the equation of this least squares line. Use the template below to write your answer. Round the values of the intercept and slope to four significant figures. 2 marks

☐ = ☐ + ☐ × ☐

e. The number of megapixels needed to construct an accurate electronic image of an oyster is approximately normally distributed.

Measurements made on recently harvested oysters showed that:

- 97.5% of the electronic images contain less than 4.6 megapixels
- 84% of the electronic images contain more than 4.3 megapixels.

Use the 68–95–99.7% rule to determine, in megapixels, the mean and standard deviation of this normal distribution. 2 marks

mean = ☐ standard deviation = ☐

Question 2 (5 marks)

a. The following data shows the sizes of a sample of 20 oysters rated as small, medium or large.

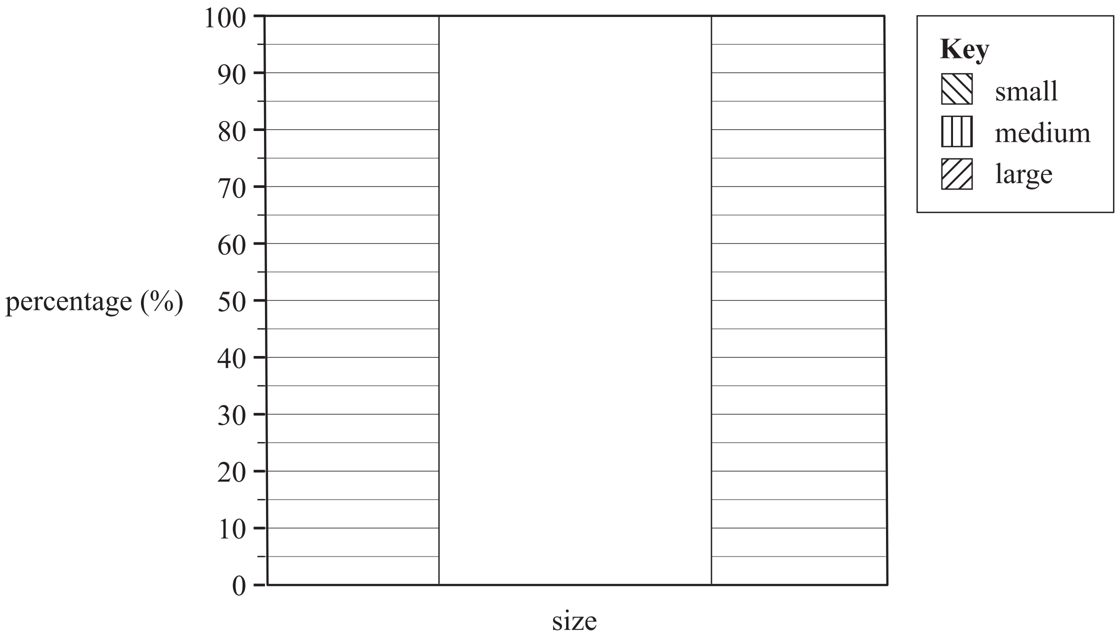
small	small	large	medium	medium
medium	large	small	medium	medium
small	medium	small	small	medium
medium	medium	medium	small	large

i. Use the data above to complete the following frequency table. 1 mark

Table 2

	Frequency	
Size	**Number**	**Percentage (%)**
small		35
medium		50
large		15
Total		100

ii. Use the percentages in Table 2 to construct a percentage segmented bar chart below. A key has been provided. 1 mark



Question 2 – continued

DO NOT WRITE IN THIS AREA

An oyster farmer has two farms, A and B.

She takes a random sample of oysters from each of the farms and has the oysters classified as small, medium or large.

The number of oysters of each size is displayed in the two-way table below.

Table 3

Oyster size	**Farm A**	**Farm B**
small	42	114
medium	124	160
large	44	46
Total	210	320

b. **i.** Calculate the percentage of the total number of oysters graded as 'large' in this investigation. Round the percentage to the nearest whole number. 1 mark

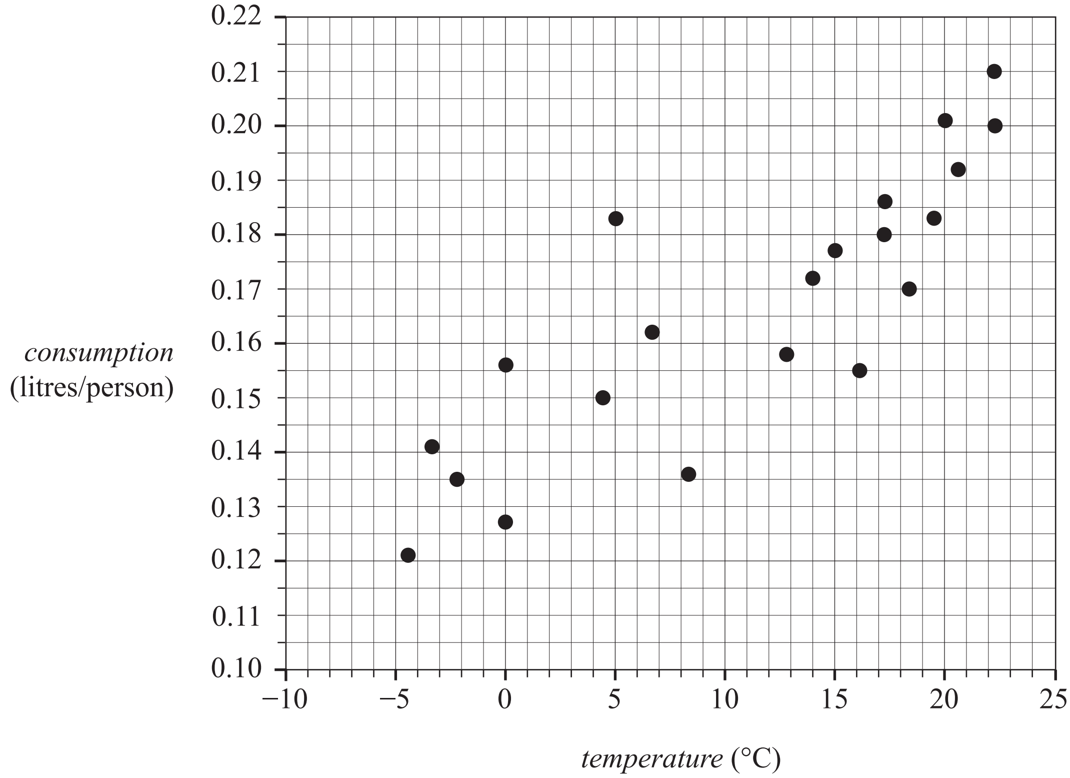
ii. The farmer believes that farm A has a greater capacity to grow larger oysters than farm B.

Does the information in Table 3 support the farmer's belief? Explain your conclusion by comparing the values of two appropriate percentages.

Round these percentages to the nearest whole number. 2 marks

Question 3 (6 marks)

The scatterplot below plots the average monthly ice cream *consumption*, in litres/person, against average monthly *temperature*, in °C. The data for the graph was recorded in the Northern Hemisphere.



When a least squares line is fitted to the scatterplot, the equation is found to be:

$$\textit{consumption} = 0.1404 + 0.0024 \times \textit{temperature}$$

The coefficient of determination is 0.7212

a. Draw the least squares line on the scatterplot graph above. 1 mark

b. Determine the value of the correlation coefficient *r*.
Round your answer to three decimal places. 1 mark

c. Describe the association between average monthly ice cream *consumption* and average monthly *temperature* in terms of strength, direction and form. 1 mark

strength	
direction	
form	

DO NOT WRITE IN THIS AREA

Question 3 – continued

d. Referring to the equation of the least squares line, interpret the value of the intercept in terms of the variables *consumption* and *temperature*. 1 mark

e. Use the equation of the least squares line to predict the average monthly ice cream *consumption*, in litres per person, when the monthly average *temperature* is −6 °C. 1 mark

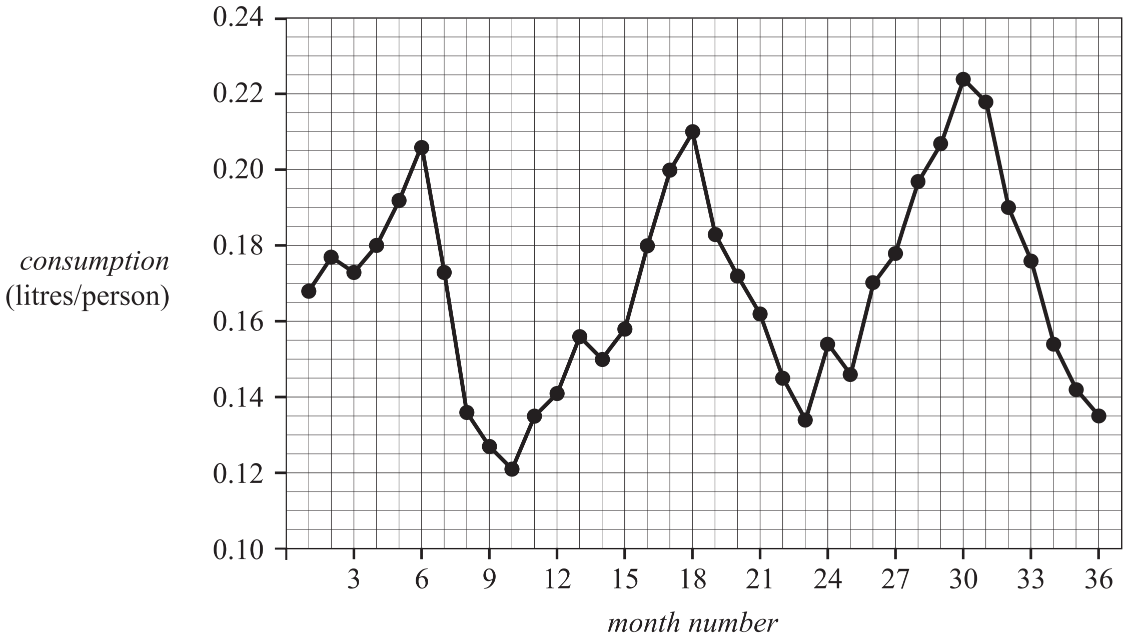
f. Write down whether this prediction is an interpolation or an extrapolation. 1 mark

Question 4 (4 marks)

The time series plot below shows the average monthly ice cream *consumption* recorded over three years, from January 2010 to December 2012.

The data for the graph was recorded in the Northern Hemisphere.

In this graph, month number 1 is January 2010, month number 2 is February 2010 and so on.



a. Identify a feature of this plot that is consistent with this time series having a seasonal component. 1 mark

b. The long-term seasonal index for April is 1.05.

Determine the deseasonalised value for average monthly ice cream *consumption* in April 2010 (month 4).

Round your answer to two decimal places. 1 mark

Question 4 – continued

c. Table 4 below shows the average monthly ice cream *consumption* for 2011.

Table 4

Consumption **(litres/person)**												
Year	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2011	0.156	0.150	0.158	0.180	0.200	0.210	0.183	0.172	0.162	0.145	0.134	0.154

Show that, when rounded to two decimal places, the seasonal index for July 2011 estimated from this data is 1.10. 2 marks

DO NOT WRITE IN THIS AREA

Instructions

- Answer **all** questions in the spaces provided.
- Write your responses in English.
- In all questions where a numerical answer is required, you should only round your answer when instructed to do so.
- Unless otherwise indicated, the diagrams in this book are **not** drawn to scale.

Data analysis

Question 1 (8 marks)

Table 1 lists the Olympic year, *year*, and the gold medal-winning height for the men's high jump, *Mgold*, in metres, for each Olympic Games held from 1928 to 2020. No Olympic Games were held in 1940 or 1944, and the 2020 Olympic Games were held in 2021.

Table 1

year	*Mgold* (m)
1928	1.94
1932	1.97
1936	2.03
1948	1.98
1952	2.04
1956	2.12
1960	2.16
1964	2.18
1968	2.24
1972	2.23
1976	2.25
1980	2.36
1984	2.35
1988	2.38
1992	2.34
1996	2.39
2000	2.35
2004	2.36
2008	2.36
2012	2.33
2016	2.38
2020	2.37

Data: <www.olympics.com/en/olympic-games/olympic-results>

a. For the data in Table 1, determine:

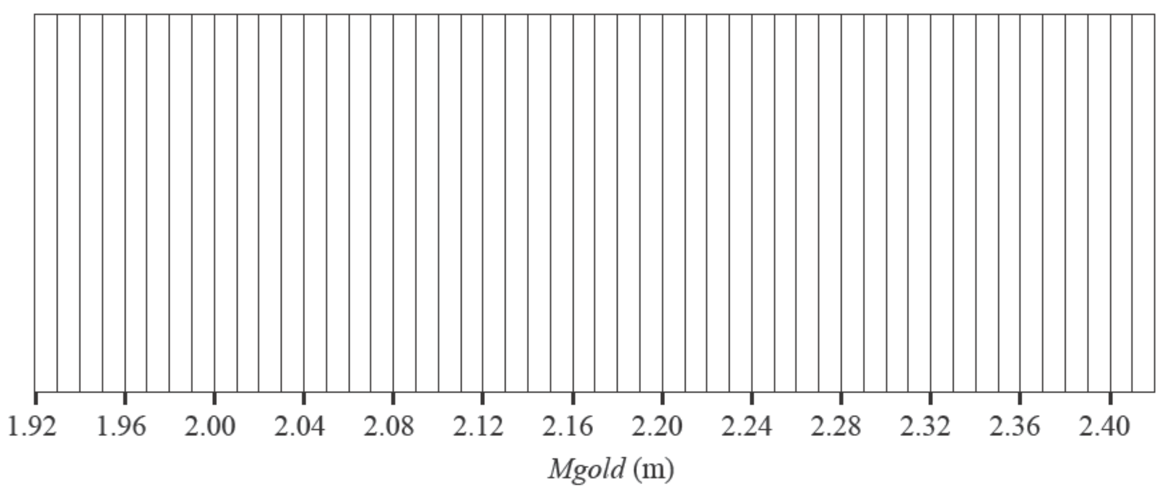
i. the maximum *Mgold* in metres 1 mark

ii. the percentage of *Mgold* values greater than 2.25 m. 1 mark

b. The mean of these *Mgold* values is 2.23 m, and the standard deviation is 0.15 m.

Calculate the standardised z-score for the 2000 *Mgold* of 2.35 m. 1 mark

c. Construct a boxplot for the *Mgold* data in Table 1 on the grid below. 2 marks



Question 1 continues on the next page.

d. A least squares line can also be used to model the association between *Mgold* and *year*.

Using the data from Table 1, determine the equation of the least squares line for this data set.

Use the template below to write your answer.

Round the values of the intercept and slope to three significant figures. 2 marks

$Mgold = \boxed{} + \boxed{} \times year$

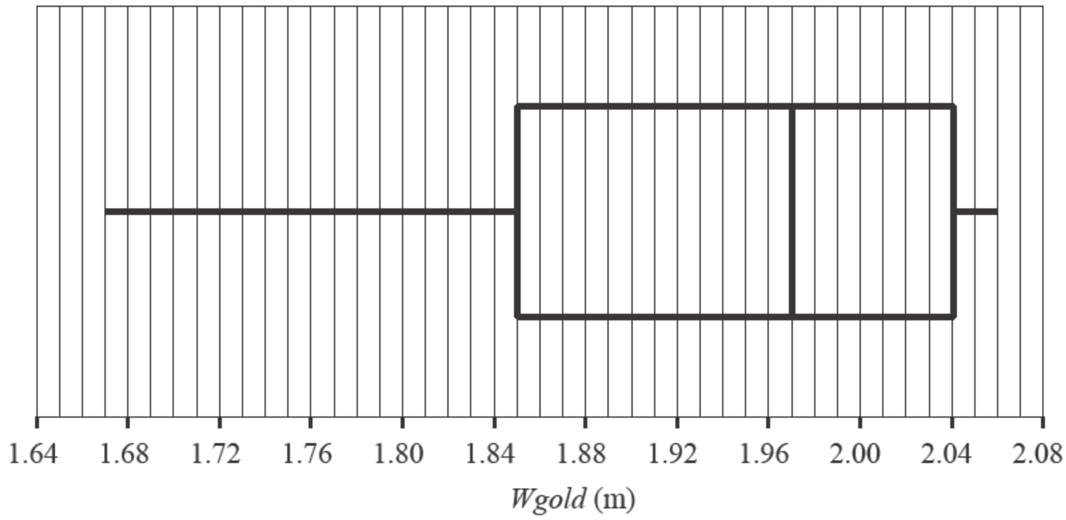
e. The coefficient of determination is 0.857

Interpret the coefficient of determination in terms of *Mgold* and *year*. 1 mark

Do not write in this area.

Question 2 (4 marks)

The boxplot below displays the distribution of all gold medal-winning heights for the women's high jump, *Wgold*, in metres, for the 19 Olympic Games held from 1948 to 2020.



Data: Adapted from <www.olympics.com/en/olympic-games/olympic-results>

a. Describe the shape of this data distribution. 1 mark

b. For this boxplot, what is the **smallest possible** number of *Wgold* heights lower than 1.85 m? 1 mark

c. **i.** Using the boxplot, show that the lower fence is 1.565 m and the upper fence is 2.325 m. 1 mark

ii. Referring to the boxplot, the lower fence and the upper fence, explain why no outliers exist. 1 mark

Question 3 (10 marks)

The Olympic gold medal-winning height for the women's high jump, *Wgold,* is often lower than the best height achieved in other international women's high jump competitions in that same year.

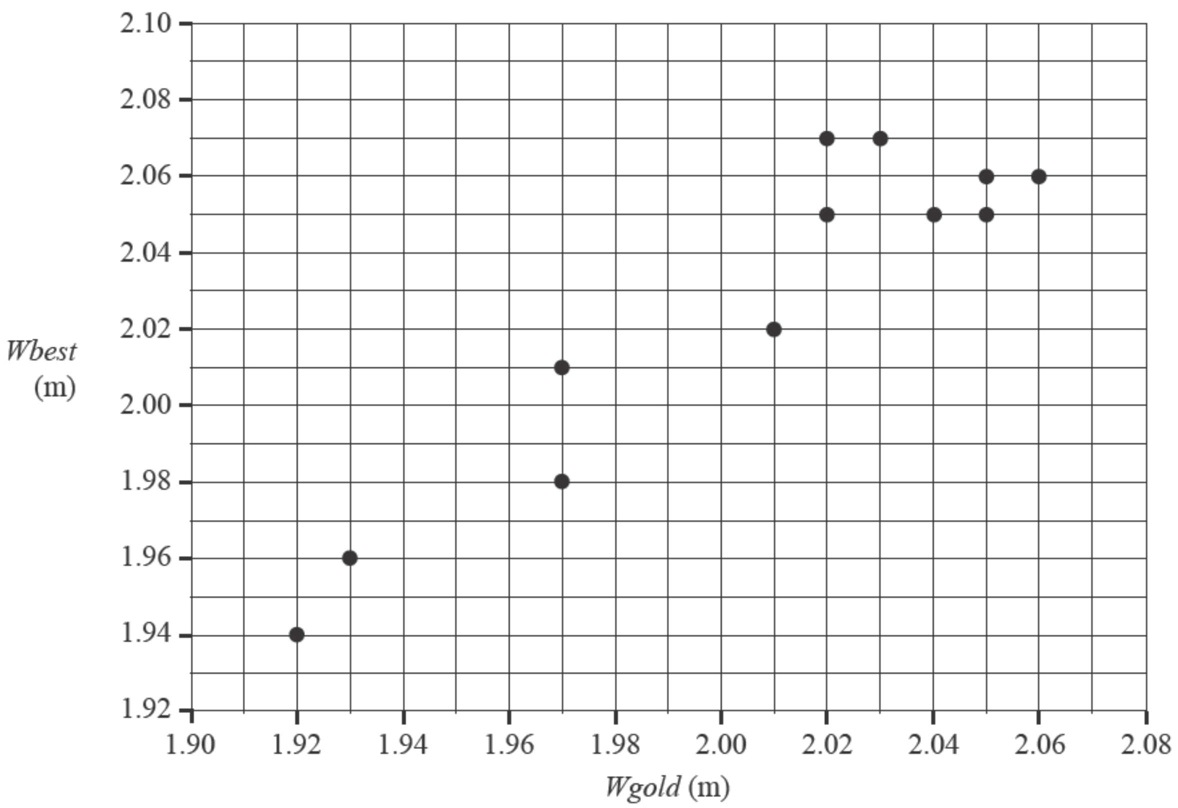
Table 2 lists the Olympic year, *year*, the gold medal-winning height, *Wgold*, in metres, and the best height achieved in all international women's high jump competitions in that same year, *Wbest*, in metres, for each Olympic year from 1972 to 2020.

A scatterplot of *Wbest* versus *Wgold* for this data is also provided.

Table 2

year	1972	1976	1980	1984	1988	1992	1996	2000	2004	2008	2012	2016	2020
Wgold (m)	1.92	1.93	1.97	2.02	2.03	2.02	2.05	2.01	2.06	2.05	2.05	1.97	2.04
Wbest (m)	1.94	1.96	1.98	2.07	2.07	2.05	2.05	2.02	2.06	2.06	2.05	2.01	2.05

Data: <www.olympics.com/en/olympic-games/olympic-results>;
<www.worldathletics.org/records/all-time-toplists/jumps/high-jump/outdoor/women/senior>



Do not write in this area.

When a least squares line is fitted to the scatterplot, the equation is found to be:

$$Wbest = 0.300 + 0.860 \times Wgold$$

The correlation coefficient is 0.9318

a. Name the response variable in this equation. 1 mark

____

b. Draw the least squares line on the scatterplot on **page 6**. 1 mark

c. Determine the value of the coefficient of determination as a percentage.
Round your answer to one decimal place. 1 mark

____

d. Describe the association between *Wbest* and *Wgold* in terms of strength and direction. 1 mark

strength	
direction	

e. Referring to the equation of the least squares line, interpret the value of the slope in terms of the variables *Wbest* and *Wgold*. 1 mark

____

____

____

f. In 1984, the *Wbest* value was 2.07 m for a *Wgold* value of 2.02 m.

Show that when this least squares line is fitted to the scatterplot, the residual value for this point is 0.0328 2 marks

____

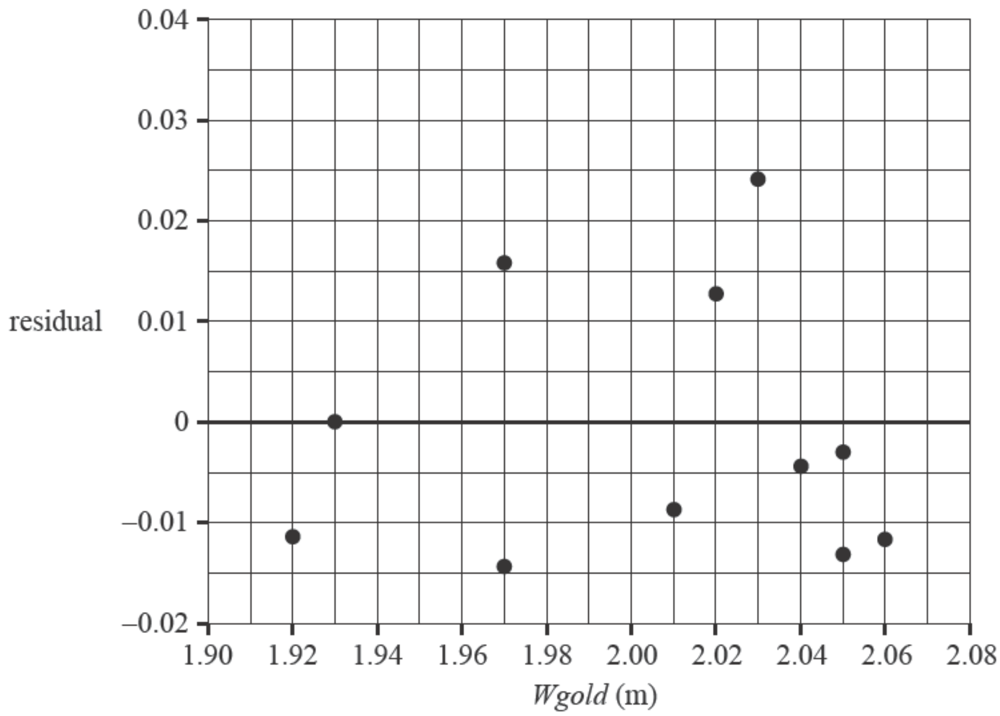
____

____

____

Question 3 continues on the next page.

g. The residual plot obtained when the least squares line was fitted to the data is shown below. The residual value from **part f** is missing from the residual plot.



i. Complete the residual plot by adding the residual value from **part f**, drawn as a cross (X), to the residual plot above. 1 mark

ii. In **part b**, a least squares line was fitted to the scatterplot.

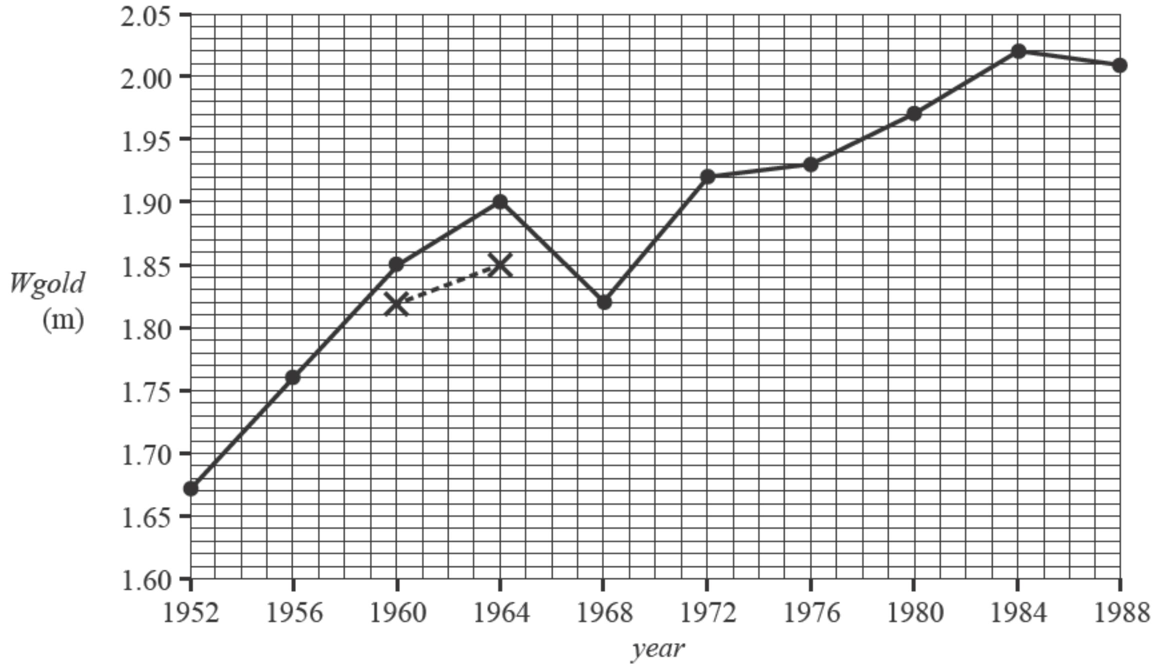
Does the residual plot from **part g** justify this? Briefly explain your answer. 1 mark

h. In 1964, the gold medal-winning height, *Wgold*, was 1.90 m. When the least squares line is used to predict *Wbest*, it is found to be 1.934 m.

Explain why this prediction is not likely to be reliable. 1 mark

Question 4 (2 marks)

The time series plot below shows the gold medal-winning height for the women's high jump, *Wgold*, in metres, for each Olympic year, *year*, from 1952 to 1988.



Data: <www.olympics.com/en/olympic-games/olympic-results>

A five-median smoothing process will be used to smooth the time series plot above.

The first two points have been placed on the graph with crosses (X) and joined by a dashed line (---).

a. Complete the five-median smoothing by marking smoothed values with crosses (X) joined by a dashed line (---) on the time series plot above. 1 mark

b. Identify **two** qualitative features that best describe the time series plot above. 1 mark

Instructions

- Answer **all** questions in the spaces provided.
- Write your responses in English.
- In all questions where a numerical answer is required, you should only round your answer when instructed to do so.
- Unless otherwise indicated, the diagrams in this book are **not** drawn to scale.

Data analysis

Question 1 (8 marks)

Data was collected to investigate the behaviour of tides in Sydney Harbour.

There are usually two high tides and two low tides each day.

The variables in this study were:

- Day: the day number in the sample
- LLT: the height of the lowest low tide for that day (in metres)
- HHT: the height of the highest high tide for that day (in metres)

Table 1 displays the data collected for a sample of 14 consecutive days in February 2021.

Table 1

Day	***LLT* (m)**	***HHT* (m)**
1	0.43	1.65
2	0.49	1.55
3	0.55	1.44
4	0.61	1.42
5	0.68	1.42
6	0.73	1.42
7	0.72	1.42
8	0.65	1.47
9	0.57	1.55
10	0.48	1.64
11	0.39	1.74
12	0.30	1.83
13	0.25	1.90
14	0.22	1.92

Data based on: http://www.bom.gov.au/australia/tides/

a. For the HHT values in Table 1:

i. Calculate the mean, in metres.

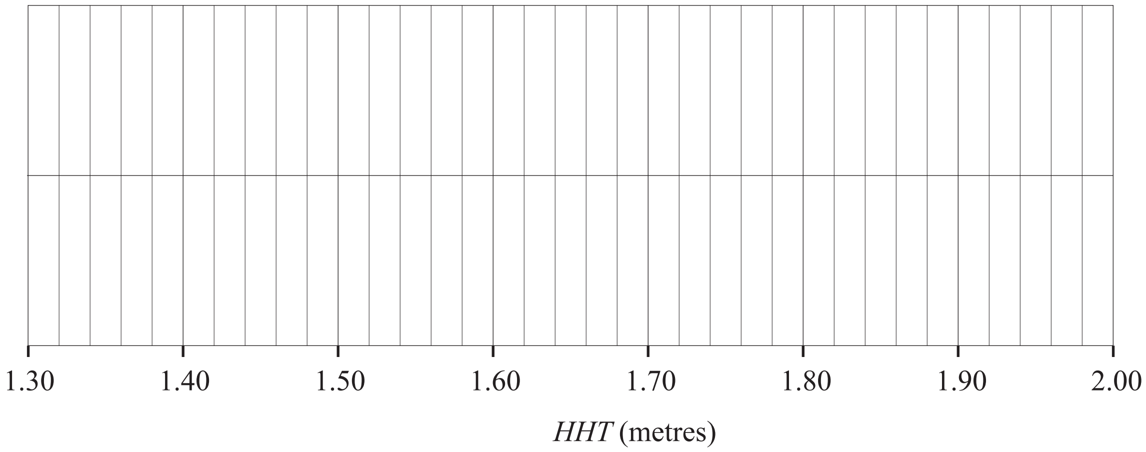
Round your answer to one decimal place. 1 mark

ii. Calculate the standard deviation, in metres.

Round your answer to three decimal places. 1 mark

---

b. Use the *HHT* data from Table 1 to construct a boxplot on the grid below. 2 marks



c. The five-number summary of the *LLT* data is shown in Table 2 below.

Table 2

Minimum	Q1	Median	Q3	Maximum
0.22	0.39	0.52	0.65	0.73

Show that the minimum *LLT* value of 0.22 m is **not** an outlier. 2 marks

---

---

d. A least squares line can be used to model the association between *LLT* and *HHT*. In this model, *HHT* is the response variable.

Use the data from Table 1 to determine the equation of this least squares line.

Round the values of the intercept and slope to four significant figures.

Write your answers in the boxes provided. 2 marks

[ ] = [ ] + [ ] × [ ]

Question 2 (4 marks)

In the year 2021 there were 12 days where a new moon appeared in Sydney Harbour.

For these 12 days the occurrence of the highest high tides (HHT) and lowest low tides (LLT) occurred before, on or after the appearance of the new moon.

These records are displayed in the table below.

Table 3

LLT	on	before	before	after	after	after
	after	after	after	after	after	after

HHT	on	before	before	before	after	after
	after	on	on	after	after	after

Data: http://www.bom.gov.au/australia/tides/

a. Use the data from Table 3 to complete the following frequency table. 2 marks

Table 4

	Frequency	
Occurrence	LLT	HHT
before		
on		
after		
Total	12	12

b. Does the data in Table 4 support the contention that the occurrence of the LLT is associated with the appearance of the new moon?

Explain your conclusion by comparing the values of appropriate percentages. 2 marks

Do not write in this area.

Question 3 (2 marks)

The time difference between successive high tides and low tides is approximately normally distributed.

Analysis of the 2021 tide chart showed that

- 99.85% of the time differences are more than 4.88 hours
- 16% of the time differences are less than 5.76 hours.

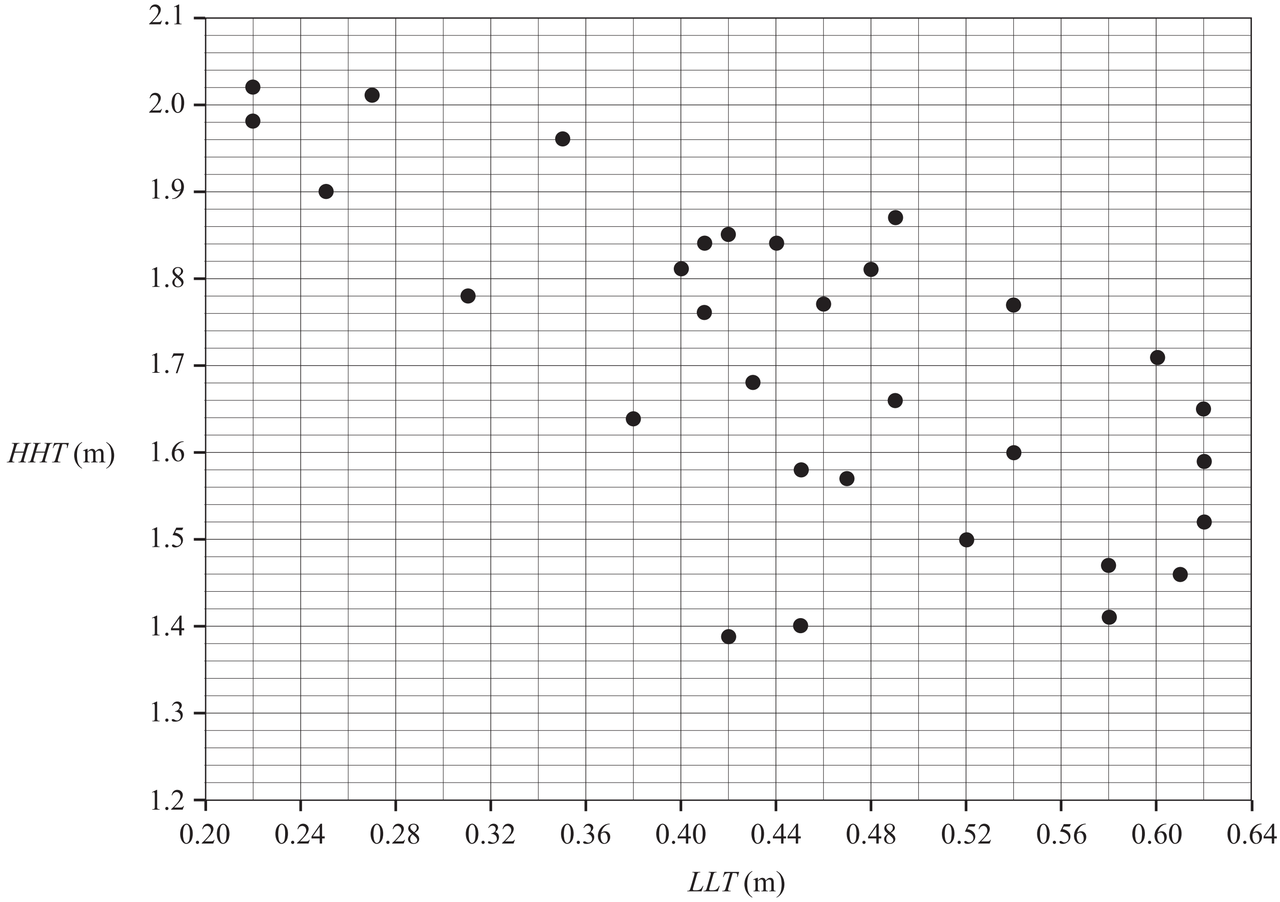
Use the 68-95-99.7% rule to determine the mean and standard deviation for this normal distribution.

mean = ☐ standard deviation = ☐

Question 4 (7 marks)

In another study, the heights, in metres, of the highest high tide (HHT) for that day and the lowest low tide (LLT) for that day were recorded in Sydney Harbour for the 31 days of July 2021.

A scatterplot of this data is shown below.



Data: http://www.bom.gov.au/australia/tides/

When a least squares line is fitted to the scatterplot, the equation is found to be:

$$HHT = 2.19 - 1.08 \times LLT$$

The coefficient of determination is 0.4709

a. Draw the graph of the least squares line on the **scatterplot above**. 1 mark

(*Answer on the scatterplot above*.)

b. Determine the value of the correlation coefficient r.

Round your answer to three decimal places. 1 mark

c. Describe the association between HHT and LLT in terms of form and direction. 1 mark

form	
direction	

d. Interpret the slope of the least squares line in terms of the variables HHT and LLT. 1 mark

e. In this investigation, the HHT value was 1.81 m for an LLT value of 0.40 m.

Show that when this least squares line is fitted to the scatterplot, the residual for this point is 0.052 2 marks

f. The mean of the HHT values for July 2021 is 1.70 m.

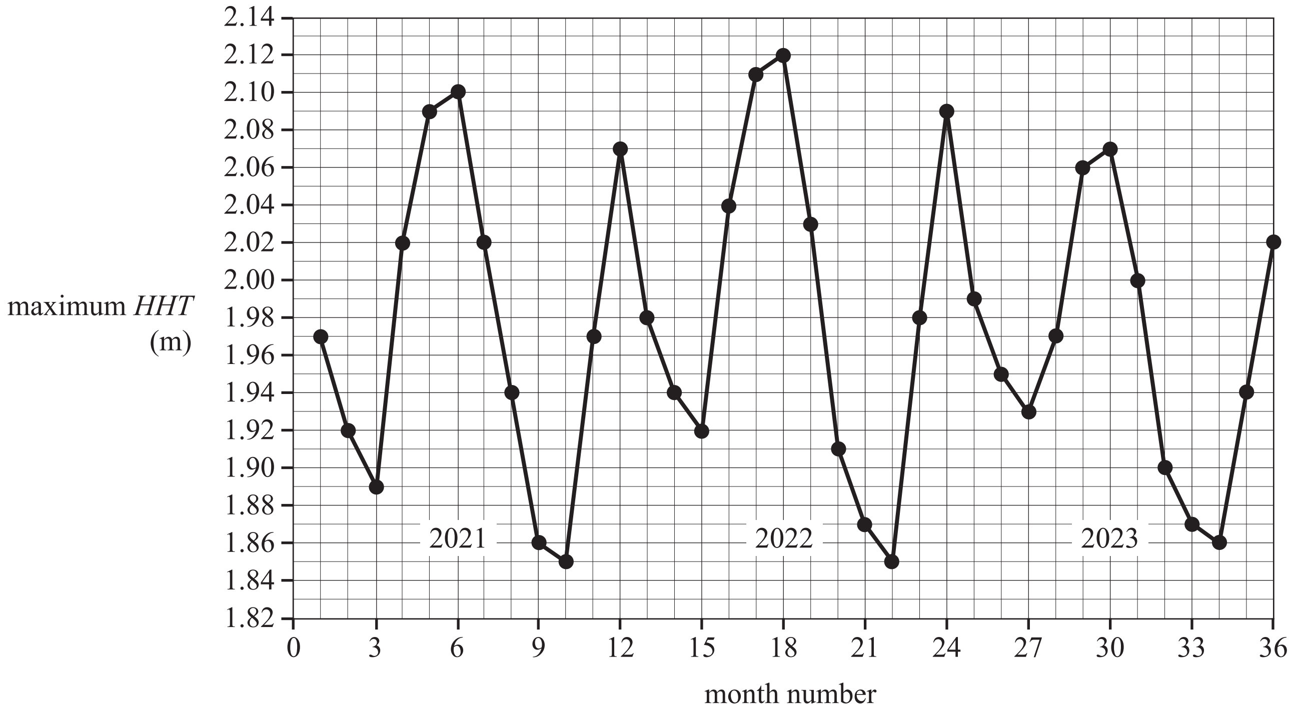
Calculate the mean of the LLT values.

Round your answer to two decimal places. 1 mark

Question 5 (2 marks)

The height, in metres, of the maximum highest high tides (HHT) for Sydney Harbour change from month to month during the year. This is shown in the time series plot below for the years 2021, 2022 and 2023.

In this graph, month number 1 is January 2021, month number 2 is February 2021, and so on.



Data based on: http://www.bom.gov.au/australia/tides/

The average height, in metres, of the maximum HHT for each year, rounded to two decimal places, is given in the table below.

Table 5

Year	2021	2022	2023
Average height of the maximum HHT (m)	1.98	1.99	1.96

The three years of data shown in this graph and in Table 5 will be used to calculate seasonal indices.

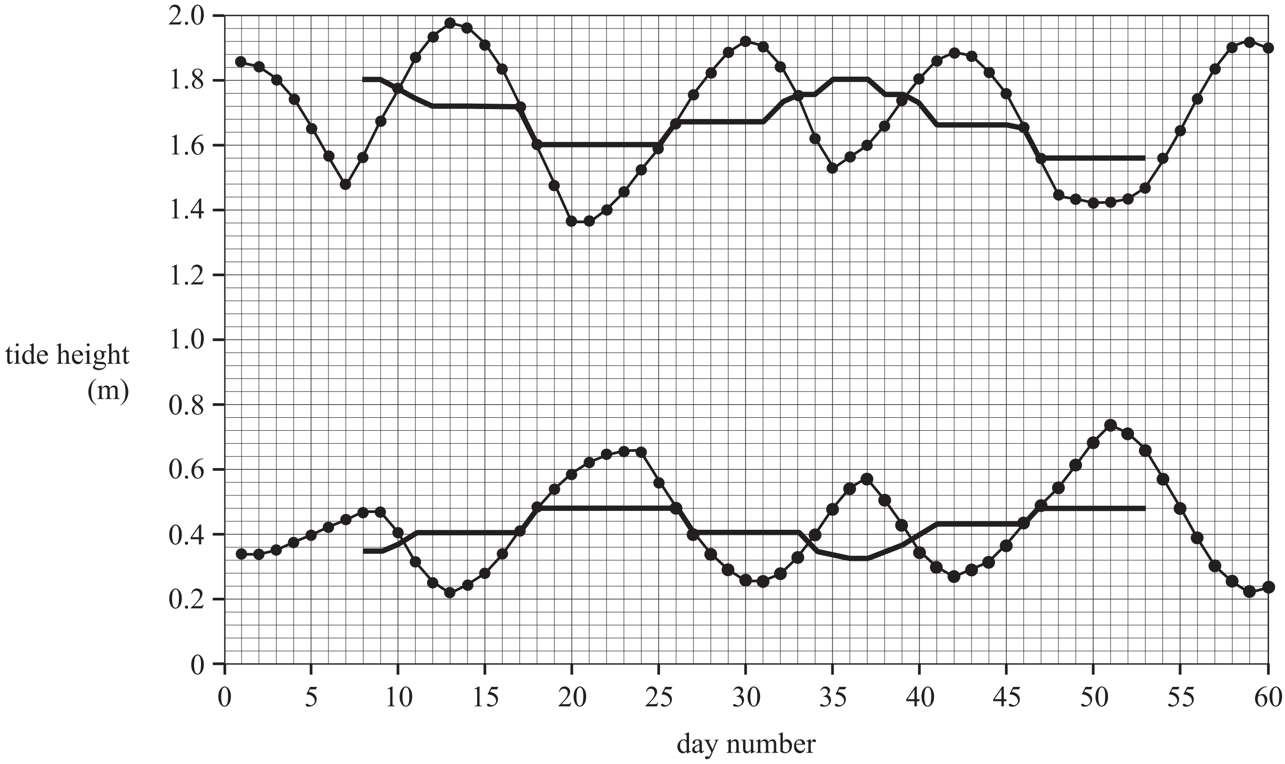
Determine the seasonal index for March.

Round your answer to two decimal places.

Question 6 (1 mark)

The time series plot below shows the height, in metres, of the highest high tides (HHT) and lowest low tides (LLT) for Sydney for the first 60 days of 2021.

The thick line for each tide type shows the results of smoothing using a moving median.



Data based on: http://www.bom.gov.au/australia/tides/

Complete the sentence below by entering a number in the box provided.

Both the HHT data and the LLT data have been smoothed using [ ]-median smoothing.