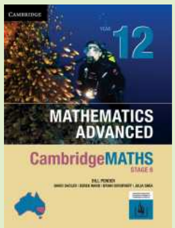
NORMANHURST BOYS HIGH SCHOOL

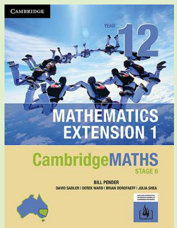
MATHEMATICS ADVANCED
MATHEMATICS EXTENSION 1/2
(YEAR 12 COURSE)

Topic summary and exercises:

(A) (X1) (X2) **Descriptive Statistics and Bivariate Data Analysis**

With references to





Name: ....

Initial version by M. Ho, with additional suggestions from H. Lam, June 2020. Last updated November 24, 2021.
Various corrections by students and members of the Mathematics Department at Normanhurst Boys High School.

Acknowledgements Pictograms in this document are a derivative of the work originally by Freepik at `http://www.flaticon.com`, used under CC BY 2.0.

Symbols used

(!) Beware! Heed warning.

(A) Mathematics Advanced content.

(X1) Mathematics Extension 1 content.

(L) Literacy: note new word/phrase.

$\mathbb{R}$ the set of real numbers

$\forall$ for all

Syllabus outcomes addressed

MA12-8 solves problems using appropriate statistical processes

Syllabus subtopics

MA-F1 Working with Functions

MA-S1 Probability and Discrete Probability Distributions

MA-S2 Descriptive Statistics and Bivariate Data Analysis

Gentle reminder

- For a thorough understanding of the topic, *every* question in this handout is to be completed!
- Additional questions from *CambridgeMATHS Year 12 Advanced* or *CambridgeMATHS Year 12 Extension 1* will be completed at the discretion of your teacher.
- Remember to copy the question into your exercise book!

Contents

Section 1

Describing Data

Learning Goal(s)

Knowledge
Differentiate between ordinal and nominal data, as well as discrete and continuous data

Skills
Calculate the mean, median and standard deviation of data sets

Understanding
Recognise that discrete data can occur in decimal increments

By the end of this section am I able to:

31.1 Classify data relating to a single random variable

31.3 Calculate measures of central tendency and spread and investigate their suitability in real-world contexts and use to compare large datasets

1.1 Ⓡ Random variables

Definition 1

A *random experiment* is an experiment with .... possible outcome.

Definition 2

A *random variable* X is the .... of a trial of a random experiment. The various outcomes of the experiment are represented by the *values* of X.

1.2 Ⓡ Types of data

There are two basic types of data: categorical and numerical.

Definition 3

Categorical data is .... and can be grouped in categories. There are two types:

1. *Ordinal data* can be logically .... in some sort of order.
2. *Nominal data* has .... order.

Definition 4

Numerical data is .... and can be counted or measured.

1. *Discrete data* has a .... number of distinct values and can only take particular values.
2. *Continuous data* has an .... number of possible values in a particular range.

Example 1

Determine whether the data sets are nominal, ordinal, discrete or continuous.

(a) The weight of fruit taken from each individual tree in an orchard.

(b) The starting letter or digit of the numberplate for each vehicle in a car park.

(c) The number of brands of clothing available in a shopping centre.

(d) The degree of support for the new jumper design for your local sporting team.

1.3 (R) Measures of central tendency

Definition 5

The *mean* .... or .... of a data set: ....

where x_i are the .... , f_i are their .... , and n is the .... of the sample.

Definition 6

The *median* .... (second quartile) is the .... score of the data set when arranged in .... order.

- For an odd number of scores, the median is the .... score.
- For an even number of scores, the median is the .... of the two .... scores

Definition 7

The *mode* is the score with the ....

- A *unimodal* data set has one .... mode
- A data set with two or more modes is .... , ... or ....

Example 2

[**2013 General HSC Q14**] The July sale prices for properties in a suburb were:

552000, 595000, 607000, 607000, 682000 and 685000.

On 1 August, another property in the same suburb was sold for over one million dollars.

If this property had been sold in July, what effect would it have had on the mean and median sale prices for July?

(A) Both the mean and the median would have changed.

(B) Neither the mean nor the median would have changed.

(C) The mean would have changed and the median would have stayed the same.

(D) The mean would have stayed the same and the median would have changed.

Further exercises

(A) **Ex 9A**

- Q1-2, 7-9, 12-15

(X1) **Ex 15A**

- Q1-2, 7-9, 12-15

1.4 Ⓡ Measures of spread

The *interquartile range* is also a measure of spread, which will be covered later.

Definition 8

The *range* of a data set is the difference between the .... and .... scores.

Definition 9

The *variance* ...... of a data set is the average of the .... - .... of the scores from the mean:

.... or ....

where x_i are the ...., f_i are their ...., and n is the total number of ....

Definition 10

The *standard deviation* .... is a measure of the typical spread of scores from the mean.

The standard deviation of a data set is the .... of the variance:

....

Important note

The standard deviation has the same units as the .... in the data set, while the variance has the square units of the ....

Example 3

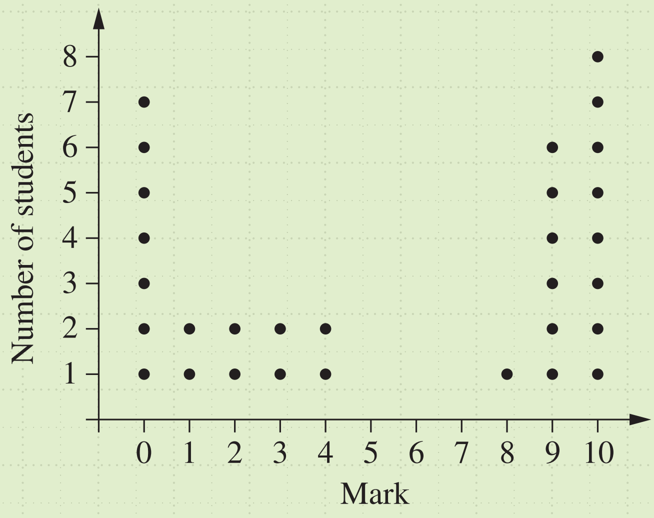
[**2014 General 2 HSC Q30**] The expenditures per primary school student for 15 countries are:

5.9, 7, 7.6, 8.4, 11.2, 11.2, 13.7, 17.1, 18.7, 21.1, 22, 22.5, 23.2, 24.9, 27.6

Calculate the mean, $\overline{x}$, and the standard deviation, σ, of the data, to two decimal places.

Example 4

[**2017 General 2 HSC Q29**] All the students in a class of 30 did a test. The marks, out of 10, are shown in the dot plot.



(i) Find the median mark.

(ii) The mean test mark is 5.4. The standard deviation of the test marks is 4.22. Using the dot plot, calculate the percentage of the marks which lie within one standard deviation of the mean.

Further exercises

Ⓐ **Ex 9B**

- Q2-3

Ⓧ1 **Ex 15B**

- Q2-3

Section 2

Univariate Data

Learning Goal(s)

Knowledge
Interpret and construct tabular and graphical displays of data

Skills
Calculate measures of data using a NESA approved calculator to compare data displays

Understanding
Regonise that Pareto charts accentuate the data values with most significant impact

By the end of this section am I able to:

31.2 Organise, interpret and display data into appropriate tabular and/or graphical representations including Pareto charts, cumulative frequency distribution tables or graphs, parallel box-plots and two-way tables

31.4 Summarise and interpret grouped and ungrouped data through appropriate graphs and summary statistics

31.5 Identify outliers and investigate and describe the effect of outliers on summary statistics

31.6 Describe, compare and interpret the distributions of graphical displays and/or numerical datasets and report findings in a systematic and concise manner.

2.1 (R) Frequency tables

Definition 11

A *frequency table* summarises data and shows frequency values .... for individual or grouped data.

Important note

Grouped numerical data in frequency distribution tables may represent discrete or continuous scores.

Definition 12

Grouping data organising scores into intervals of .... to provide a clearer overview.

Important note

Grouping data involves ignoring information, which results in its summary statistics to be an approximation of that of the raw data

Definition 13

The *class centre* (abbreviated to ....) is the midpoint of each interval used in the grouping.

Definition 14

The *cumulative frequency* (abbreviated to ....) is the number of scores that are .... or .... a given score, for numerical data.

Important note

A frequency distribution table can be extended to a cumulative frequency distribution table by taking the accumulating sums of the frequencies.

Example 5

For each of the following frequency, distribution tables, classify the data as discrete or continuous:

(a)

Number of CDs	f
0 – 9	3
10 – 19	5
20 – 29	6
30 – 39	2
40 – 49	1

(b)

Distance travelled (km)	f
$0- < 1$	3
$1- < 2$	5
$2- < 3$	6
$3- < 4$	2
$4- < 5$	1

Example 6

[2016 General 2 HSC Q21] A grouped data frequency table is shown.

Class interval	Frequency
1 – 5	3
6 – 10	6
11 – 15	8
16 – 20	9

(A) 6.5 (B) 10.5 (C) 11.9 (D) 12.4

What is the mean for this set of data?

Example 7

A bowler records the number of wickets he takes in his past 20 cricket matches:

Score (x)	f	fx	$c.f.$
0	3	0	3
1	6	6	9
2	2	4	11
3	7	21	18
4	2	8	20
Total	20	39	

If he bowls 2 wickets in his next match, which measure will change?

(A) mode (B) median (C) mean (D) range

Further exercises

Ⓐ **Ex 9A**
- Q3-5

Ⓐ **Ex 9B**
- Q1, 4

(x1) **Ex 15A**
- Q3-5

(x1) **Ex 15B**
- Q1, 4

2.1.1 Calculating measures using technology

Important note

All instructions for calculator functions detailed are based off the CASIO fx-82AU PLUS II

Steps

To enter a data set into a frequency distribution table on a CASIO calculator:

1. Frequency column setting: `AC` , `SHIFT` `MODE` , `↓` , `3` for `STAT`:

- `1` for `ON` (max. 20 rows)
- `2` for `OFF` (max. 40 rows)

2. Enter `STAT` mode: `AC` , `MODE` , `2` for `STAT` , `1` for `1-VAR`

3. Fill in the frequency distribution table: use the arrow keys to navigate the cells , use the number keys and `=` to enter data values

Steps

To edit the frequency distribution table:

- `AC` , `SHIFT` `1` , `2` for `STAT` , `1` for `1-VAR Data`

To reset the frequency distribution table:

(i) Enter `STAT` mode again , `AC` , `MODE` , `2` for `STAT` , `1` for `1-VAR`

OR

(ii) In the frequency table screen: `SHIFT` `1` , `3` for `Edit` , `2` for `Del-A`

Steps

Calculating 'Variance': AC , SHIFT 1 , 4 for `Var`, then:

- *mean*: 2 for $\overline{x}$, =
- *standard deviation*: 3 for σ`x` , =
- *size*: 1 for `n` , =

Steps

Calculating 'Sum': AC , SHIFT 1 , 5 for `Sum`, then:

- *sum of scores*: 2 for Σx , =
- *sum of squared scores*: press 1 for Σx^2 , =

Steps

Calculating 'MinMax': AC , SHIFT 1 , 5 for `MinMax`, then:

- *minimum*: 1 for `minX` , =
- *maximum*: 2 for `maxX` , =
- *lower quartile*: 3 for Q_1 , =
- *median*: 4 for `med` , =
- *upper quartile*: 5 for Q_3 , =

Important note

The median, lower quartile and upper quartile functions are only available on *CASIO fx-82 PLUS II* model calculators and above.

2.2 Ⓡ Frequency histograms and frequency polygons

Definition 15

A *frequency histogram* is visually similar to a .... graph, but with no .... in between that columns.

The subintervals on the horizontal axis of frequency histograms are often called *bins*.

Important note

When constructing a frequency histogram:

- The first column is usually placed one half-column width from the vertical axis.
- The columns join up with no gaps.
- Each column is centred on the value for individual data.
- Each column is centred on the class centre for grouped data.
- A *scale break* or zigzag on the x or y-axis indicates that the data displayed does not include all the values that exist on the number line used.

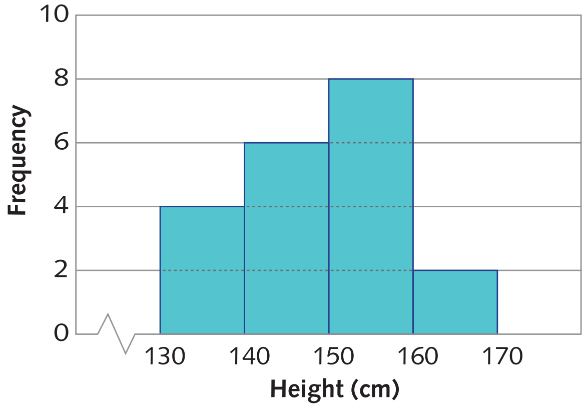
Definition 16

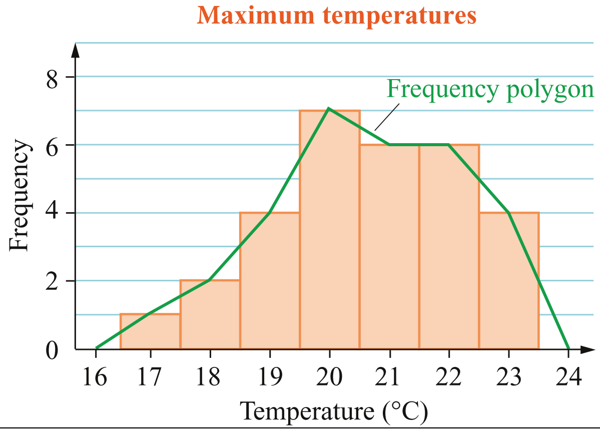
A *frequency polygon* is a .... often constructed on the same graph as a frequency histogram.

Important note

When constructing a frequency polygon:

- The plotted points are at the centre of the top of each column of the histogram.
- The plotted points are joined with intervals.
- The polygon starts on the left, on the horizontal axis at the previous value or class centre.
- The polygon ends on the right, on the horizontal axis at the next value or class centre.





Important note

Coarser grouping in frequency histograms is more practical, as too many columns can make the data display difficult to interpret.

Example 8

At the start of Year 7, Cedar Heights High School gave 40 students a spelling test marked out of 10. The raw results were organised into a frequency table.

Mark (x)	1	2	3	4	5	6	7	8	9	10
Frequency (f)	2	4	2	1	6	8	7	6	2	2

(a) Draw a histogram and frequency polygon for the original data.

(b) Construct a frequency distribution table by grouping the data into subintervals of 1-2, 3-4, ... , including a row for the class centres.

(c) Draw a histogram and frequency distribution polygon for the grouped data.

(d) Compare and comment on what the two data displays have shown.

Definition 17

A *cumulative frequency histogram* is formed by stacking the columns of a frequency histogram.

Important note

The height of the last column of a cumulative frequency histogram is the .... of the sample.

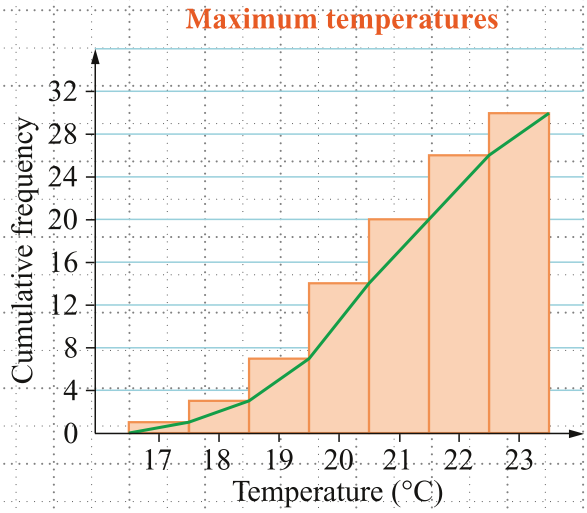
Definition 18

A *cumulative frequency polygon* or *ogive* shows .... and so is drawn slightly differently.

Important note

When constructing a cumulative frequency polygon:

- The polygon starts at zero, at the bottom left-hand corner of the first column (no scores have yet been accumulated).
- The polygon passess through the top right-hand corner of each column (scores less than or equal to the upper bound of the class interval are plotted).
- The polygon finishes at the top right-hand corner of the last column (its height equals the total size of the sample),



Example 9

Returning to the spelling test marks from Example 8:

Mark (x)	1	2	3	4	5	6	7	8	9	10
Frequency (f)	2	4	2	1	6	8	7	6	2	2

(a) Insert a cumulative frequency row and construct a cumulative frequency histogram and ogive for the data.

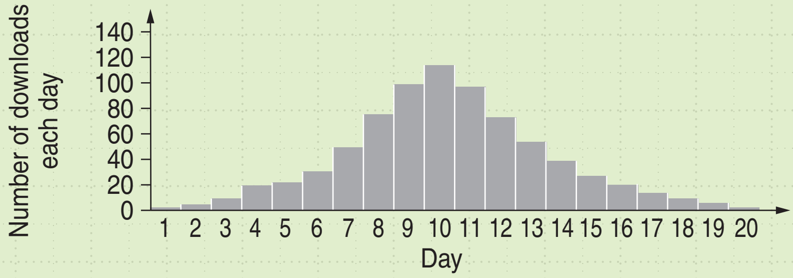
(b) Group the data by pairing the marks and construct a grouped frequency distribution table, including a cumulative frequency row.

(c) Construct a cumulative frequency histogram and ogive for the grouped data.

(d) Find the median of the original data and the grouped data, and compare them.

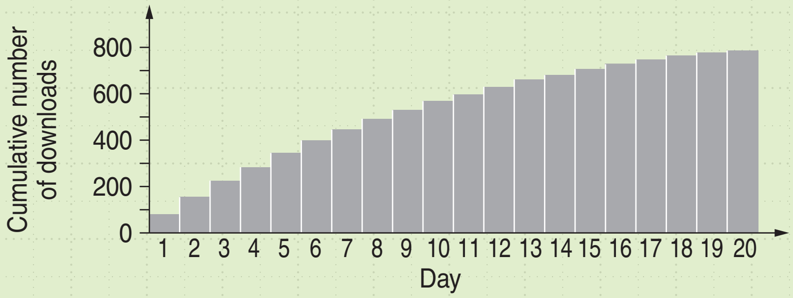
Example 10

[2021 Adv HSC Q4] The number of downloads of a song on each of twenty consecutive days is shown in the following graph.

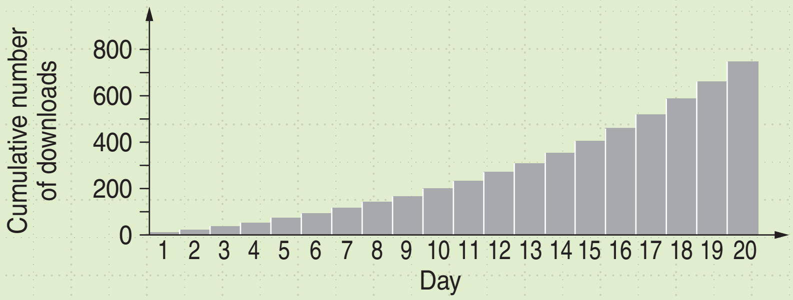


Which of the following graphs best shows the cumulative number of downloads up to and including each day?

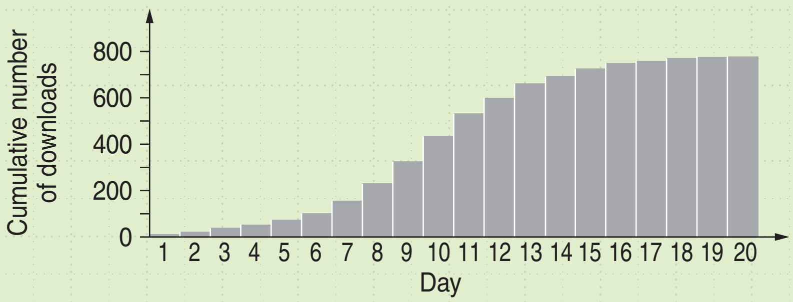
A.



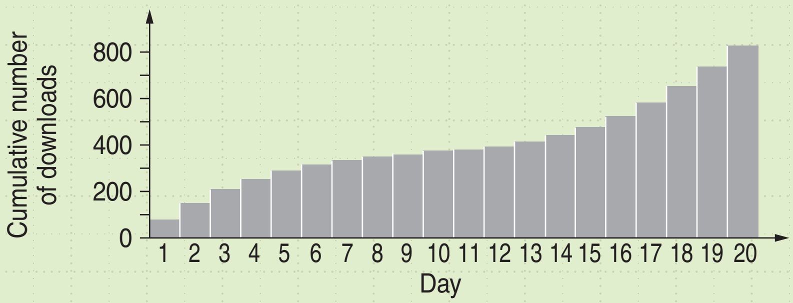
B.



C.



D.

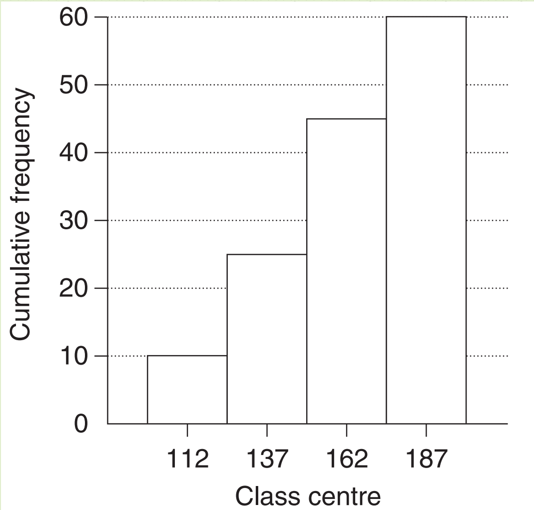


Example 11

[2010 General HSC Q26] A new shopping centre has opened near a primary school. A survey is conducted to determine the number of motor vehicles that pass the school each afternoon between 2:30 pm and 4:00 pm.

The results for 60 days have been recorded in the table and are displayed in the cumulative frequency histogram.

Score	Class centre	Frequency	Cumulative frequency
100 – 124	112	10	10
125 – 149	137	×	25
150 – 174	162	20	45
175 – 199	187	15	60



(i) Find the value of × in the table.

(ii) Carefully copy the cumulative frequency histogram and draw a cumulative frequency polygon (ogive) for this data.

(iii) Use your graph to determine the median. Show, by drawing lines on your graph, how you arrived at your answer.

(iv) Prior to the opening of the new shopping centre, the median number of motor vehicles passing the school between 2 : 30 pm and 4 : 00 pm was 57 vehicles per day.
What problem could arise from the change in the median number of motor vehicles passing the school before and after the opening of the new shopping centre? Briefly recommend a solution to this problem.

Further exercises

(A) Ex 9B
- Q5-6

(X1) Ex 15B
- Q5-6

2.3 Ⓡ Distributions of data

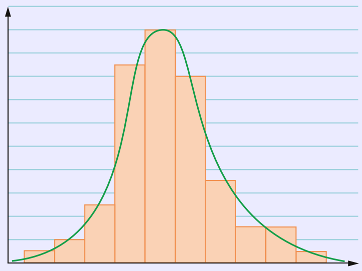
The shape of a statistical display indicates the distribution of the data.

Definition 19

Clustering occurs when the scores are .... or 'bunched up'.

Definition 20

A *symmetrical distribution* has scores balanced or .... about the centre of the distribution. This distribution is also ....

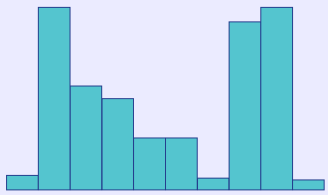


Important note

Distributions with scores that are approximately evenly spread about either side of the centre of distribution, are still classified as symmetrical distributions.

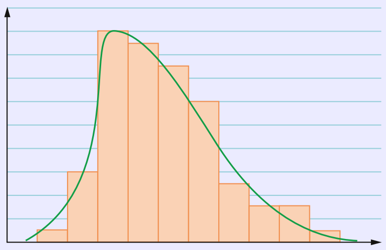
Definition 21

The distribution is .... as it has .... modes.

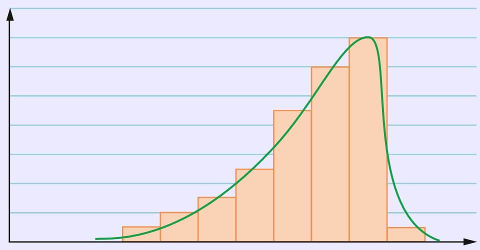


Definition 22

A data distribution is *skewed* in the direction where the data is .... There is a long .... in this direction of the distribution. The data is .... in the opposite direction of the skew.



- .... skewed, or skewed to the ....
- Most of the scores are relatively ....



- .... skewed, or skewed to the ....
- Most of the scores are relatively ....

2.4 Pareto charts

Definition 23

- The *Pareto chart* consists of a column graph with columns arranged in .... order and a .... percentage line graph, drawn together on the same chart.

Fill in the spaces

- The chart usually has .... vertical axes: .... on the left and .... on the right.
- The purpose of the chart is to identify the .... problems, particularly in quality control.

The line graph indicates the added .... of each problem.

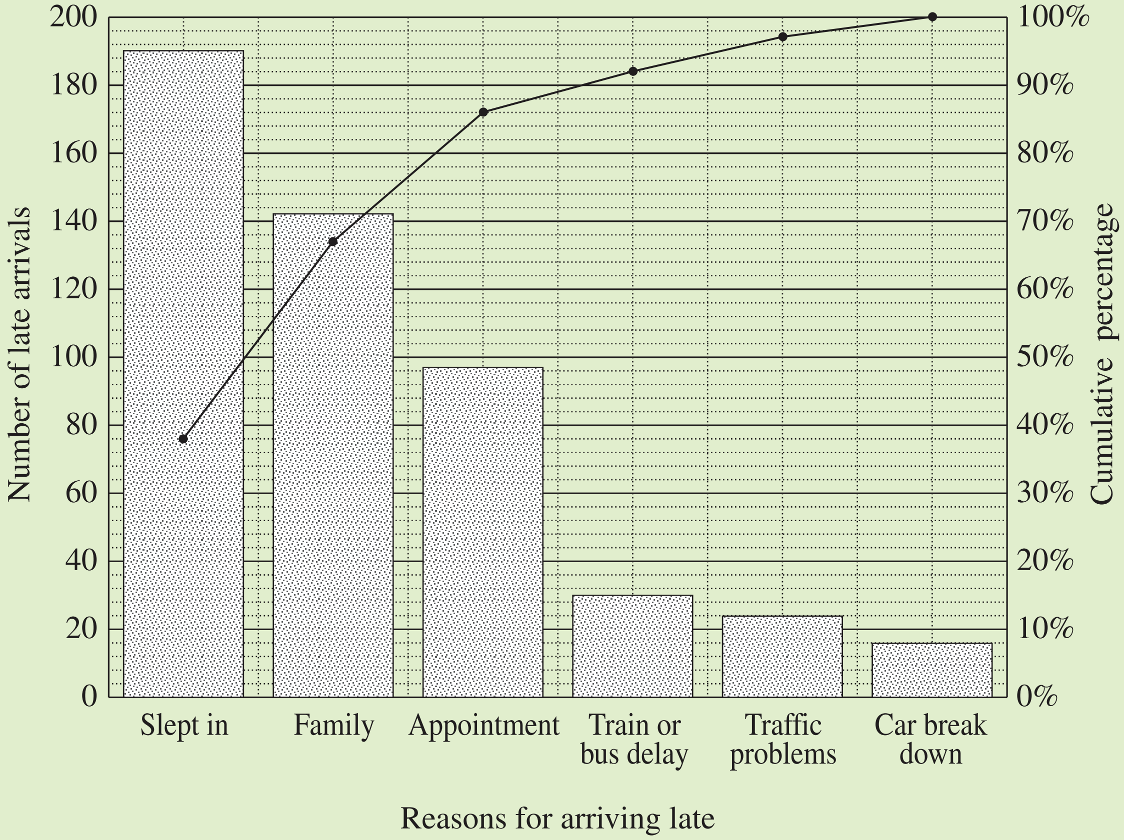
- If the line graph rises steeply and then levels out, the first two or three problems have the ....
- If the line graph rises at a steady rate, all problems have roughly .... and so the columns will be at ....

Important note

- The column graph is plotted using the frequency column of the frequency table, arranged in descending order.
- The cumulative line graph is plotted using a cumulative percentage column.

Example 12

[2020 Advanced HSC Sample Q5] A school collected data on the reasons given by students for arriving late. The Pareto chart shows the data collected.



What percentage of students gave the reason 'Train or bus delay'?

(A) 6% (B) 15% (C) 30% (D) 92%

Example 13

An online seller of clothing summarised the complaints received in a month in the following table.

Type of complaint	**No. of complaints**	**Percentage**	**Cumulative Percentage**
Problems completing the order online	50		
Difficulty accessing website	30		
Cancelled order	9		
Wrong article sent	5		
Overcharging for delivery	4		
Late delivery	2		

(a) Fill in the table's *percentage* and *cumulative percentage* columns.

(b) Construct a Pareto chart for this information.

(c) Which problems account for 80% of the complaints?

Example 14

A bank wants to identify the most significant problems it experiences in the delay of processing credit card applications. A random survey was conducted and the results are shown in the table.

Cause of delay	Frequency (f)
Incorrect address	36
Can't read	17
No signature	73
Wrong from	11
Other	3

(a) Arrange the problems by frequency in descending order and insert a cumulative percentage column.

(b) Construct a Pareto chart of the survey data.

(c) What area(s) should the manager concentrate on improving?

(d) What percentage of her problem does this solve?

Further exercises

(A) Ex 9A

- Q9-11

(X1) Ex 15A

- Q9-11

Exercises Source: (?, ?, Ex 7G Q6-9)

PROBLEM SOLVING, REASONING AND JUSTIFICATION

6 A random survey of customer complaints yielded the data shown on the right.

a Draw a Pareto chart to illustrate this information.

b What area(s) should this company concentrate on improving?

c What percentage of the problems would this solve?

Type of complaint	Frequency
Packaging	11
Delivery	7
Invoice wrong	36
Product quality	22
Other	4

7 The owner of a shoe store takes a random sample of customer complaints. The results are shown in the table on the right.

a Draw a Pareto chart to illustrate this information.

b Before the survey, the manager thought that it was the limited range of styles being offered that was the main reason for the decline in her business and she blamed the supplier. What percentage of the problem was caused by limited styles?

c If you were the shoe store owner, what area(s) would you concentrate on improving?

d What percentage of the problems would your improvements from part **c** solve?

Type of complaint	Frequency
Difficult parking	77
Salesperson rude	9
Poor lighting	5
Layout confusing	8
Limited sizes	37
Limited styles	11
Other	3

8 Pareto charts can be drawn using a spreadsheet. After entering the data as you would for a table, go to: **Insert** > **Chart** > **Histogram** > **Pareto.**

a Use your computer spreadsheet to draw Pareto charts for questions **5–7**.

b What is different about the charts drawn by Excel compared to your hand-drawn charts?

CHALLENGE

9 A restaurant manager takes a random survey of customer complaints in order to increase the patronage of his restaurant. The results are shown in the table on the right.

a Draw a Pareto chart to illustrate this information.

b If you were the manager, what area(s) would you concentrate on improving?

c What percentage of the problems would your improvements from part **b** solve?

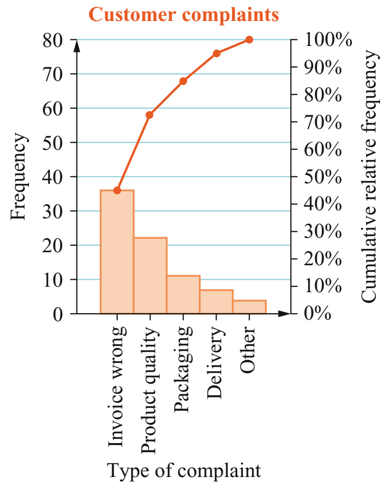
Type of complaint	Frequency
Rude staff	6
No atmosphere	8
Small portions	17
Too noisy	19
Too expensive	93
Limited menu	5
Dirty washrooms	8
Long delays in serving	35
Cramped seating	9

ANSWERS

6 a

Type of complaint	Frequency	Relative frequency (%)	Cumulative relative frequency (%)
Invoice wrong	36	45	45
Product quality	22	27.5	72.5
Packaging	11	13.8	86.3
Delivery	7	8.7	95
Other	4	5	100
Total	80		

ANSWERS

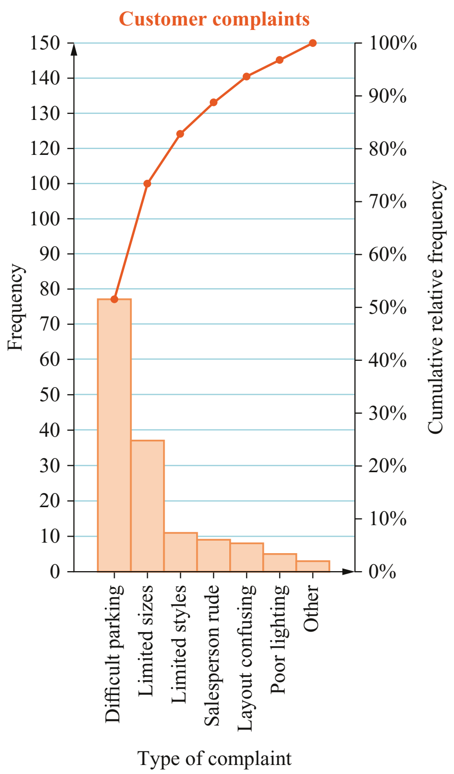


b Check invoice, check product quality, (ensure correct packaging)

c 72.5% (86.3% if packaging included)

7 a

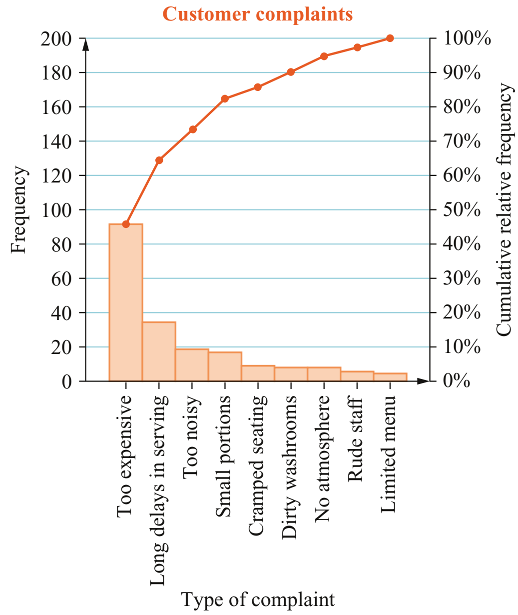
Type of complaint	Frequency	Relative frequency (%)	Cumulative relative frequency (%)
Difficult parking	77	51	51
Limited sizes	37	25	76
Limited styles	11	7	83
Salesperson rude	9	6	89
Layout confusing	8	5	94
Poor lighting	5	3	97
Other	3	2	100
Total	150		



b 7% **c** Parking, range of sizes **d** 76%

9 a

Type of complaint	Frequency	Relative frequency (%)	Cumulative relative frequency (%)
Too expensive	93	46.5	46.5
Long delays in serving	35	17.5	64
Too noisy	19	9.5	73.5
Small portions	17	8.5	82
Cramped seating	9	4.5	86.5
Dirty washrooms	8	4	90.5
No atmosphere	8	4	94.5
Rude staff	6	3	97.5
Limited menu	5	2.5	100
Total	200		



b Cost of meals, delays in serving, noise, (portion size)

c 74% (82% if portion size included)

2.5 (R) Quartiles and interquartile range

Definition 24

A data set can be divided into .... parts by .... *quartiles*; the lower quartile, median and upper quartile.

The *lower quartile* ...... is the .... of the lower .... of the data set.

The *upper quartile* ...... is the .... of the upper .... of the data set.

Important note

If the data set has an odd number of scores, the median is excluded when finding the lower and upper quartiles.

Definition 25

The *interquartile range* .... measures the .... of the middle .... of the data set:

....

Important note

The interquartile range is often a better measure of the spread of the data than the range, particularly when the data set contains outliers.

Definition 26

The *five-number summary* is a useful summary of a data set:

1. ....

2. ....

3. ....

4. ....

5. ....

Example 15

[**2016 General 2 HSC Q19**] A soccer referee wrote down the number of goals scored in 9 different games during the season.

$$2, 3, 3, 3, 5, 5, 8, 9, \square$$

The last number has been omitted. The range of the data is 10.

What is the five-number summary for this data set?

Example 16

[**2018 General 2 HSC Q26**] A cumulative frequency table for a data set is shown.

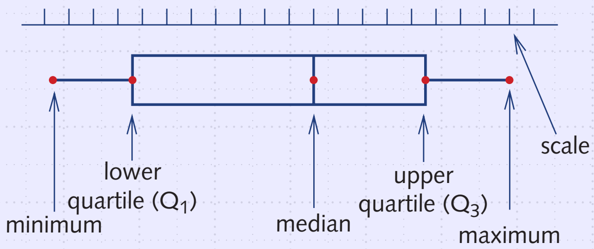
Score	1	2	3	4	5	6
Cumulative frequency	5	9	16	20	34	42

What is the interquartile range of this data set?

2.6 (R) Box plots

Definition 27

The *box-and-whisker plot* or .... displays the five-number summary.



The *box* extends from .... to the .... , and the *whiskers* extend from ...... to the .... and from ...... to the ....

Definition 28

An *outlier* is an .... data value that does not fit in with the rest of the data set and is usually .... from other data values.

Any data value more than 1.5 times the .... away from the ends of the .... is classified as an outlier:

....

....

Important note

When constructing a box plot for a data set with outliers, the whiskers end at the highest and lowest data values that lie within $1.5 \times IQR$ from the ends of the box.

The outliers are marked with dots.

Example 17

[2017 General 2 HSC Q30] A set of data has a lower quartile (Q_1) of 10 and an upper quartile (Q_3) of 16.

What is the maximum possible range for this set of data if there are no outliers?

Example 18

The waiting times in seconds at a ticket counter were as follows:

$$0, 0, 3, 5, 5, 5, 9, 10, 12, 13, 16, 17, 18, 18, 21, 22, 23, 23, 24, 24, 24, 24, 24,$$
$$25, 25, 25, 26, 26, 27, 28, 28, 28, 29, 29, 29, 30, 31, 31, 31, 32, 33, 33, 33, 33,$$
$$34, 34, 34, 34, 34, 35, 35, 35, 36, 36, 37, 38, 38, 38, 39, 39, 39, 40, 41, 41, 52$$

(a) Find Q_1, the median, Q_3 and the IQR.

(b) Find any outliers.

(c) Draw a boxplot, showing outliers.

(d) Comment on the shape of the boxplot.

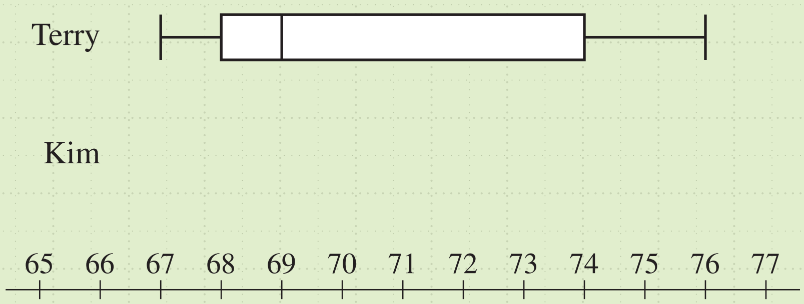
Answer: (a) $Q_1 = 22.5$, median $= 29$, $Q_3 = 34.5$, $IQR = 12$ (b) 0, 0, 3 (d) Negatively skewed

Example 19

[**2014 General 2 HSC Q29**] Terry and Kim each sat twenty class tests. Terrys results on the tests are displayed in the box-and-whisker plot shown in part (i).

(i) Kims 5-number summary for the tests is 67, 69, 71, 73, 75.

Draw a box-and-whisker plot to display Kims results below that of Terrys results.

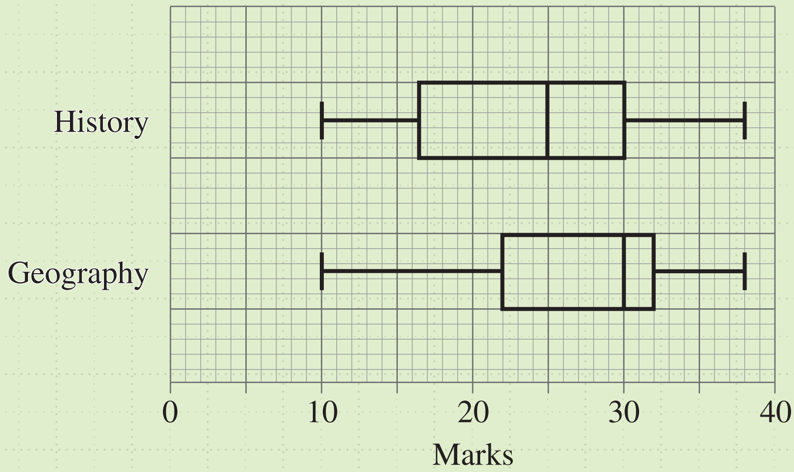


(ii) What percentage of Terry's results were below 69?

(iii) Terry claims that his results were better than Kims. Is he correct? Justify your answer by referring to the summary statistics and the skewness of the distributions.

Example 20

[2016 General 2 HSC Q22] The box-and-whisker plots show the results of a History test and a Geography test.



In History, 112 students completed the test. The number of students who scored above 30 marks was the same for the History test and the Geography test.

How many students completed the Geography test?

(A) 8 (B) 50 (C) 56 (D) 112

Example 21

The towns of Karuah and Buladelah both have a speed limit of 60 km/h. The speeds of the first 100 cars travelling through these towns, from 9 : 00 am to 10 : 00 am on Saturday, were measured and the results are shown in the table.

	Karuah	Buladelah
Mean	59	60
Median	58	58
Lower quartile	52	50
Interquartile range	18	9
Highest speed	85	105
Lowest speed	35	40

(a) Construct parallel box plots to display this data.

(b) Explain why the police might target Karuah for speeding.

(c) Explain why the police might target Buladelah for speeding.

Further exercises

(A) Ex 9C

- Q1-8

(X1) Ex 15C

- Q1-8

Section 3

Bivariate Data

Learning Goal(s)

Knowledge
Identify and describe linear association in bivariate scatterplots

Skills
Calculate Pearson's correlation coefficient and the equation of the least squares regression line using a NESA approved calculator

Understanding
Recognise the possible but not guaranteed link between correlation and causation

By the end of this section am I able to:

31.7 Construct a bivariate scatterplot to identify patterns in the data that suggest the presence of an association

31.8 Use bivariate scatterplots (constructing them where needed), to describe the patterns, features and associations of bivariate datasets, justifying any conclusions

31.9 Calculate and interpret Pearsons correlation coefficient (r) using technology to quantify the strength of a linear association of a sample

31.10 Model a linear relationship by fitting an appropriate line of best fit to a scatterplot and using it to describe and quantify associations

31.11 Use the appropriate line of best fit, both found by eye and by applying the equation of the fitted line, to make predictions by either interpolation or extrapolation

31.12 Solve problems that involve identifying, analysing and describing associations between two numeric variables

31.13 Construct, interpret and analyse scatterplots for bivariate numerical data in practical contexts

Definition 29

A *bivariate data set* consists of two different .... for each data point.

3.1 (R) Two-way tables

Definition 30

A *two-way table* or *contingency table* consists of two or more related .... combined together.

Information can be read from each ...., ... or ...., with each .... and .... representing a separate frequency table.

Example 22

[2016 General 2 HSC Q23] A group of 485 people was surveyed. The people were asked whether or not they smoke. The results are recorded in the table.

	Smokers	Non-smokers	Total
Male	88	176	264
Female	68	153	221
	156	329	485

A person is selected at random from the group.

What is the approximate probability that the person selected is a smoker OR is male?

(A) 33% (B) 18% (C) 68% (D) 87%

Example 23

[2011 General HSC Q25] At another school, students who use mobile phones were surveyed. The set of data is shown in the table.

	Pre-paid	Plan	TOTAL
Female students	172	147	319
Male students	158	103	261
TOTAL	330	250	

(i) How many students were surveyed at this school?

(ii) Of the female students surveyed, one is chosen at random. What is the probability that she uses pre-paid?

(iii) Ten new male students are surveyed and all ten are on a plan. The set of data is updated to include this information. What percentage of the male students surveyed are now on a plan? Give your answer to the nearest per cent.

Example 24

[2017 General 2 HSC Q29] A group of Year 12 students was surveyed. The students were asked whether they live in the city or the country. They were also asked if they have ever waterskied.
The results are recorded in the table.

	Have waterskied	Have never waterskied
Live in the city	150	2500
Live in the country	70	800

(i) A person is selected at random from the group surveyed.
Calculate the probability that the person lives in the city and has never waterskied.

(ii) A newspaper article claimed that Year 12 students who live in the country are more likely to have waterskied than those who live in the city.
Is this true, based on the survey results? Justify your answer with relevant calculations.

Further exercises

(A) **Ex 9A**
- Q6, 16

(X1) **Ex 15A**
- Q6, 16

3.2 Linear Association of Scatterplots

Definition 31

A *scatterplot* treats bivariate data as a series of .... to plot on a suitable set of axes.

Important note

Repeated points can be indicated with numbers inside the dot, or a code using circles, squares, crosses, larger circles, etc.

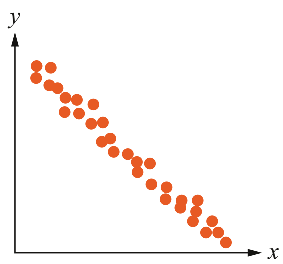
Important note

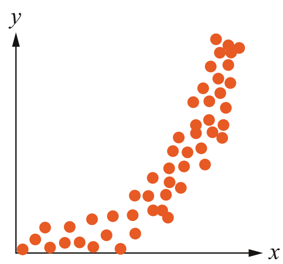
Not all bivariate data sets have clear independent and dependent variables.

Definition 32

The *correlation* of a bivariate data set measures the extent to which two variables are ....

The form of bivariate data sets can be .... or .... relationship, depending on the shape of its clustering:





Important note

Linear relationships between two variables can be described in terms of direction and strength of association.

Definition 33

The *direction* of a linear relationship can be negative or positive:

- In a *positive relationship*, one variable increases as the other variable ....
- In a *negative relationship*, one variable increases as the other variable ....

Definition 34

Pearson's correlation coefficient, ... is a statistical measure of the strength of .... (or ....) between two variables.

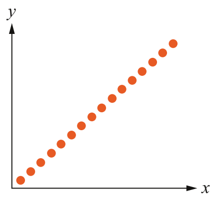
Important note

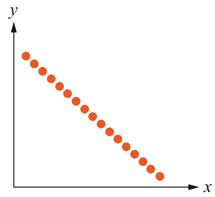
The formula for Pearson's correlation coefficient is not developed or used by hand in Mathematics Advanced/Extension 1 or Extension 2.

Laws/Results

Pearson's correlation coefficient assigns a numerical value between -1 and 1 to the correlation.

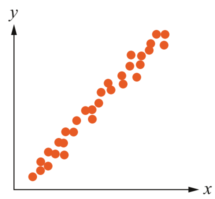
- .... *positive/negative correlation*:

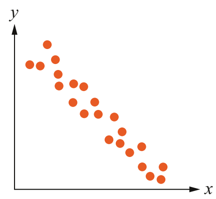




....

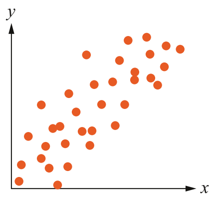
- .... *positive/negative correlation*:

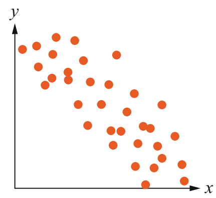




....

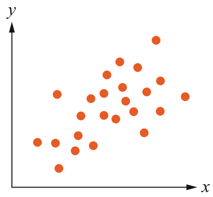
- .... *positive/negative correlation*:

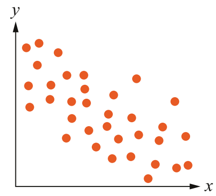




....

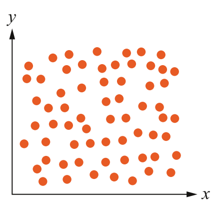
- .... *positive/negative correlation*:





....

- *No correlation*:



....

Steps

To calculate Pearson's correlation coefficient on a CASIO calculator:

1. *Bivariate frequency table*: `MODE` `2` for `STAT` , `2` for `A+BX`

2. Use the arrow and numbers keys to enter the data points into the table.

3. `AC` , `SHIFT` `1` , `5` for `Reg` , `3` for `r` , `=`

Important note

Correlation does not necessarily imply causation.

Example 25

[2012 General HSC Q11] Which of the following relationships would most likely show a negative correlation?

(A) The population of a town and the number of hospitals in that town.

(B) The hours spent training for a race and the time taken to complete the race.

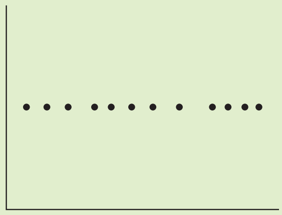
(C) The price per litre of petrol and the number of people riding bicycles to work.

(D) The number of pets per household and the number of computers per household.

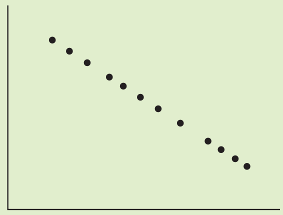
Example 26

[2017 General 2 HSC Q12] Which of the data sets graphed below has the largest positive correlation coefficient value?

A.

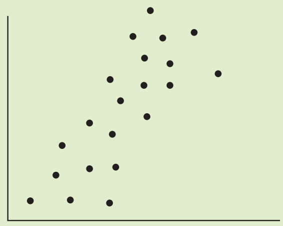


B.



C.

D.



Example 27

[2010 General HSC Q6] A survey of Year 7 students found a number of relationships with a high degree of correlation.
Which of the following relationships also demonstrates causality?

(A) Students height and the length of their arm span

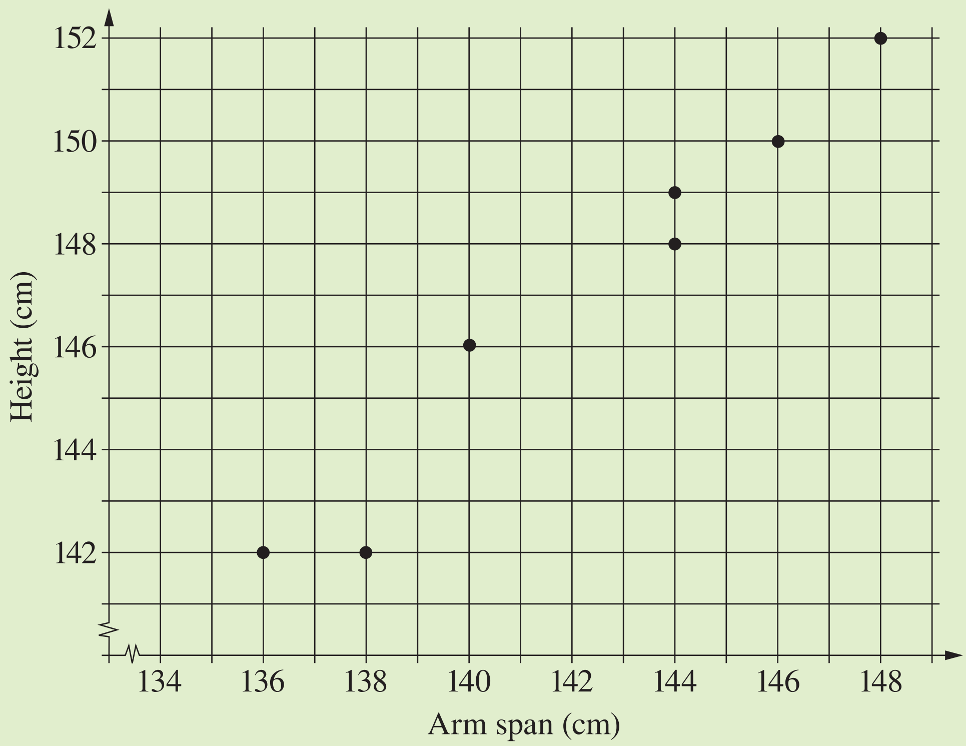
(B) The size of students left feet and the size of their right feet

(C) Students test scores in Mathematics and their test scores in Music

(D) The number of hours students spent studying for a test and their results in that test

Example 28

[2019 Standard 2 HSC Q23] A set of bivariate data is collected by measuring the height and arm span of seven children. The graph shows a scatterplot of these measurements.



(a) Calculate Pearsons correlation coefficient for the data, correct to two decimal places.

(b) Identify the direction and the strength of the linear association between height and arm span.

(c) The equation of the least-squares regression line is shown.

$$\text{Height} = 0.866 \times (\text{arm span}) + 23.7$$

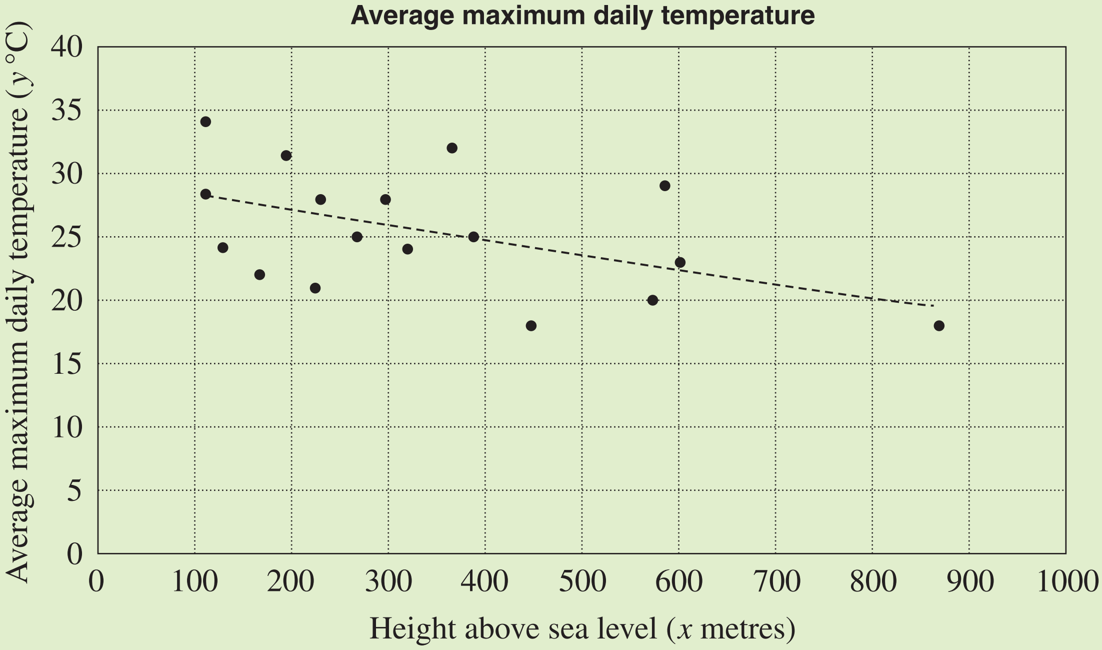
A child has an arm span of 143 cm.
Calculate the predicted height for this child using the equation of the least-squares regression line.

Answer: (a) 0.98 (b) Positive, strong (c) 147.538 cm

Example 29

[2021 Adv HSC Q17] For a sample of 17 inland towns in Australia, the height above sea level, x (metres), and the average maximum daily temperature, y (°C), were recorded.

The graph shows the data as well as a regression line.



The equation of the regression line is $y = 29.2 - 0.011x$.

The correlation coefficient is $r = -0.494$.

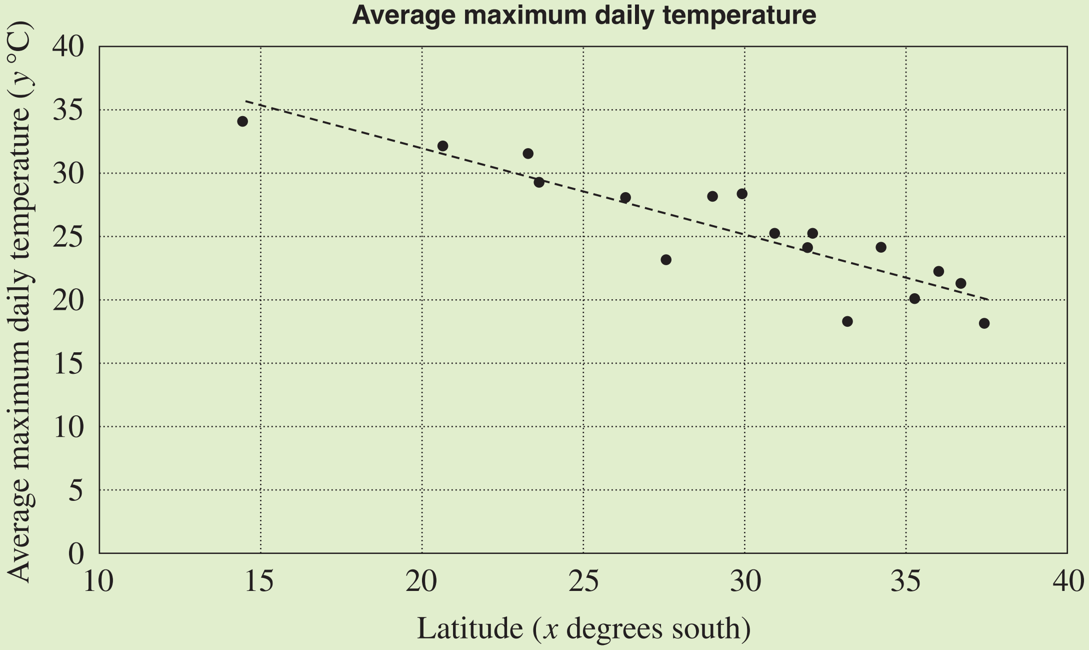
(a) (i) By using the equation of the regression line, predict the average maximum daily temperature, in degrees Celsius, for a town that is 540 m above sea level. Give your answer correct to one decimal place.

(ii) The gradient of the regression line is -0.011. Interpret the value of this gradient in the given context.

Example 29

[2021 Adv HSC Q17] *Example 29 on the preceding page continued...*

(b) The graph below shows the relationship between the latitude, x (degrees south), and the average maximum daily temperature, y (°C), for the same 17 towns, as well as a regression line.



The equation of the regression line is $y = 45.6 - 0.683x$.

The correlation coefficient is $r = -0.897$.

Another inland town in Australia is 540 m above sea level. Its latitude is 28 degrees south.

Which measurement, height above sea level or latitude, would be better to use to predict this town's average maximum daily temperature? Give a reason for your answer.

Further exercises

Ⓐ **Ex 9D**
- Q1-5

(X1) **Ex 15D**
- Q1-5

3.3 Lines of Best Fit

Definition 35

The *line of best fit* or .... is a .... line that provides a representation of all the data points in a .... that has a .... correlation.

Important note

Draw a line of best fit by eye involves fitting the line such that:

- the distance of points from the line is ....
- an approximately .... number of data points lie above and below the line

The line does not need to pass through any data points.

Definition 36

The *least-squares regression line* is a mathematically determined straight line that .... fits the data set by .... the squares of the vertical distances from the points to the line:

....

where $m =$ .... and $b =$ ....

Laws/Results

The gradient of the least-squares regression line:

....

where $r =$ Pearson's ....,
$\sigma_x =$ .... of the x-variable,
$\sigma_y =$ .... of the y-variable.

Laws/Results

The y-intercept of the least-squares regression line:

....

where $m =$ .... , $\overline{x} =$ .... of the x-variable,
and $\overline{y} =$ .... of the y-variable

Laws/Results

The equation of the least-squares regression line can also be expressed as:

. . . .

where $A =$ . . . . and $B =$ . . . .

Steps

To find the equation of the least-squares regression line on a CASIO calculator:

1. *Bivariate frequency table*:

- `MODE` `2` for `STAT` mode
- `2` for `A+BX`

2. Use the arrow and numbers keys to enter the data points into the table.

3. *y-intercept:*

- `AC` `SHIFT` `1` `5` for `Reg`
- `1` for `A`, then `=`

4. *gradient:*

- `AC` `SHIFT` `1` `5` for `Reg`
- `2` for `B`, then `=`

The line of best fit can be used to make predictions given a value of one of the two variables.

Important note

The accuracy of the predictions increases with correlation coefficients .... to $+1$ or -1, and with a .... number of data points.

Definition 37

Interpolation predicts the value of a .... data point .... the existing range of data points.

Definition 38

Extrapolation predicts the value of a .... data point .... the existing range of data points.

Important note

Extrapolation can be misleading since data relationships often do not continue without changing unpredictably.

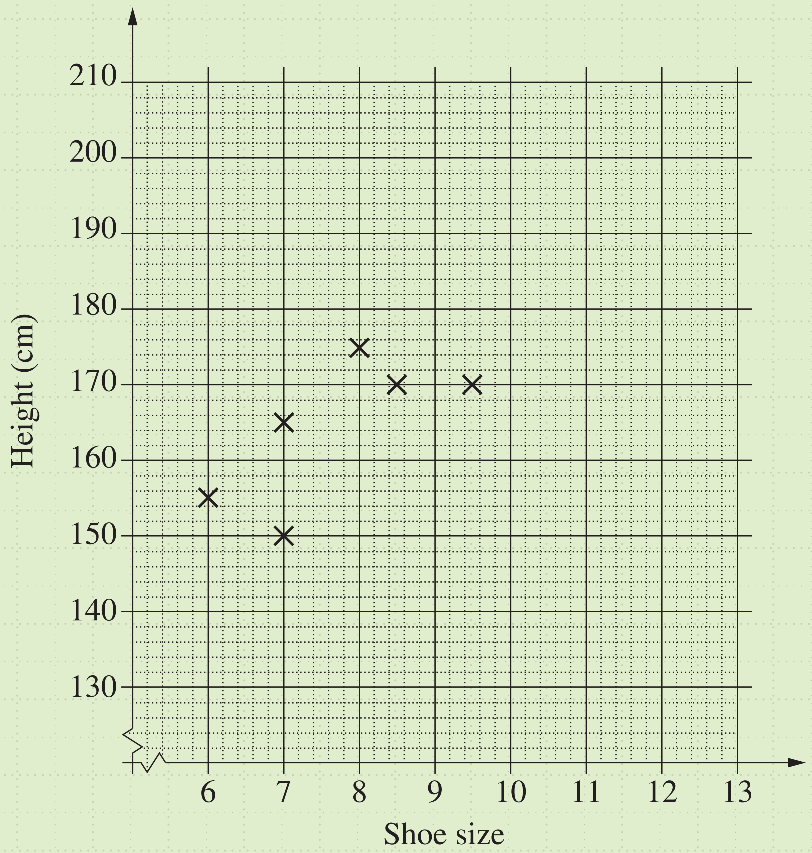
Extrapolation should not be used to predict data values far outside the range of the existing values.

Example 30

[2015 General 2 HSC Q28] The shoe size and height of ten students were recorded.

Shoe size	6	7	7	8	8.5	9.5	10	11	12	12
Height	155	150	165	175	170	170	190	185	200	195

(i) Complete the scatter plot AND draw a line of fit by eye.



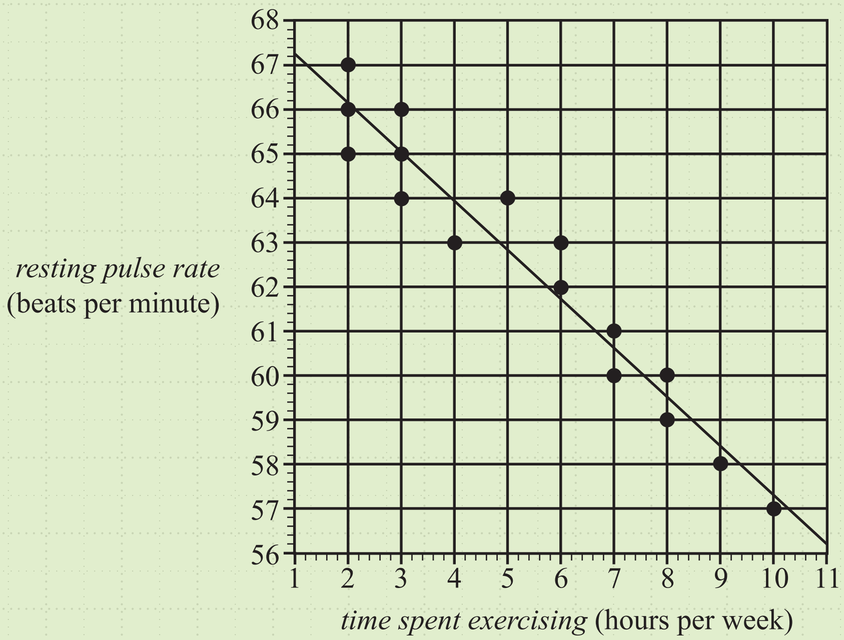
(ii) Use the line of fit to estimate the height difference between a student who wears a size 7.5 shoe and one who wears a size 9 shoe.

(iii) A student calculated the correlation coefficient to be 1 for this set of data. Explain why this cannot be correct.

Answer: (ii) 11 cm

Example 31

[2018 VCE Further Mathematics Exam 1 Q8] The scatterplot below displays the resting pulse rate, in beats per minute, and the time spent exercising, in hours per week, of 16 students. A least squares line has been fitted to the data.



The equation of this least squares line is closest to

(A) resting pulse rate $= 67.2 - 0.91\times$ time spent exercising

(B) resting pulse rate $= 67.2 - 1.10\times$ time spent exercising

(C) resting pulse rate $= 68.3 - 0.91\times$ time spent exercising

(D) resting pulse rate $= 68.3 - 1.10\times$ time spent exercising

(E) resting pulse rate $= 67.2 + 1.10\times$ time spent exercising

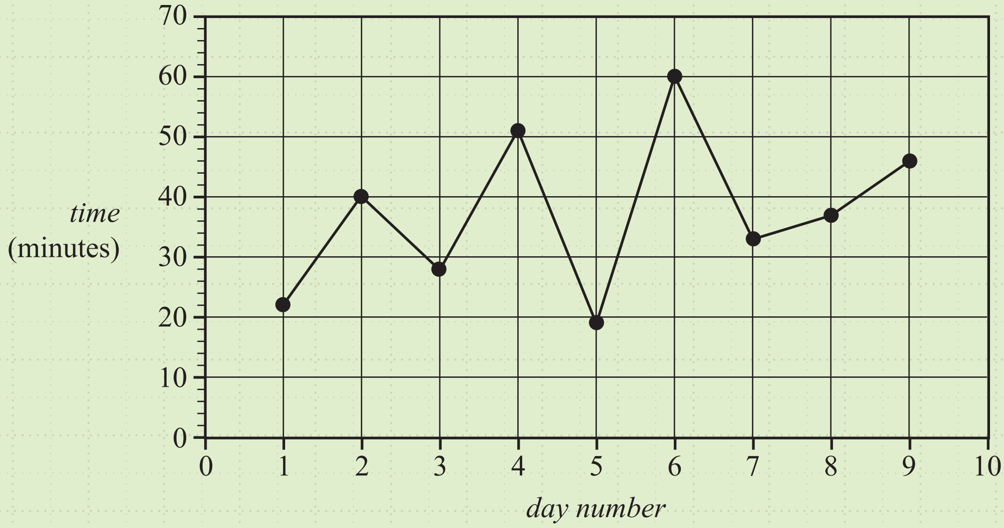
Answer: D

Example 32

[2019 VCE Further Mathematics Exam 1 Q14] The time, in minutes, that Liv ran each day was recorded for nine days.
These times are shown in the table below.

Day number	1	2	3	4	5	6	7	8	9
Time (minutes)	22	40	28	51	19	60	33	37	46

The time series plot below was generated from this data.



A least squares line is to be fitted to the time series plot shown above.
The equation of this least squares line, with day number as the explanatory variable, is closest to

(A) day number $= 23.8 + 2.29\times$ time
(B) day number $= 28.5 + 1.77\times$ time
(C) time $= 23.8 + 1.77\times$ day number
(D) time $= 23.8 + 2.29\times$ day number
(E) time $= 28.5 + 1.77\times$ day number

Example 33

[2018 VCE Further Mathematics Exam 1 Q13] The statistical analysis of a set of bivariate data involving variables x and y resulted in the information displayed in the table below.

Mean	$\overline{x} = 27.8$	$\overline{y} = 33.4$
Standard deviation	$\sigma_x = 2.33$	$\sigma_y = 3.24$
Equation of the least squares line	$y = -2.84 + 1.31x$	

Using this information, the value of the correlation coefficient r for this set of bivariate data is closest to

(A) 0.88 (B) 0.89 (C) 0.92 (D) 0.94 (E) 0.97

Answer: D

Example 34

[2019 VCE Further Mathematics Exam 1 Q11] A study was conducted to investigate the effect of drinking coffee on sleep.

In this study, the amount of sleep, in hours, and the amount of coffee drunk, in cups, on a given day were recorded for a group of adults.

The following summary statistics were generated.

	Sleep (hours)	Coffee (cups)
Mean	7.08	2.42
Standard deviation	1.12	1.56
Correlation coefficient (r)	-0.770	

On average, for each additional cup of coffee drunk, the amount of sleep

(A) decreased by 0.55 hours

(B) decreased by 0.77 hours

(C) decreased by 1.1 hours

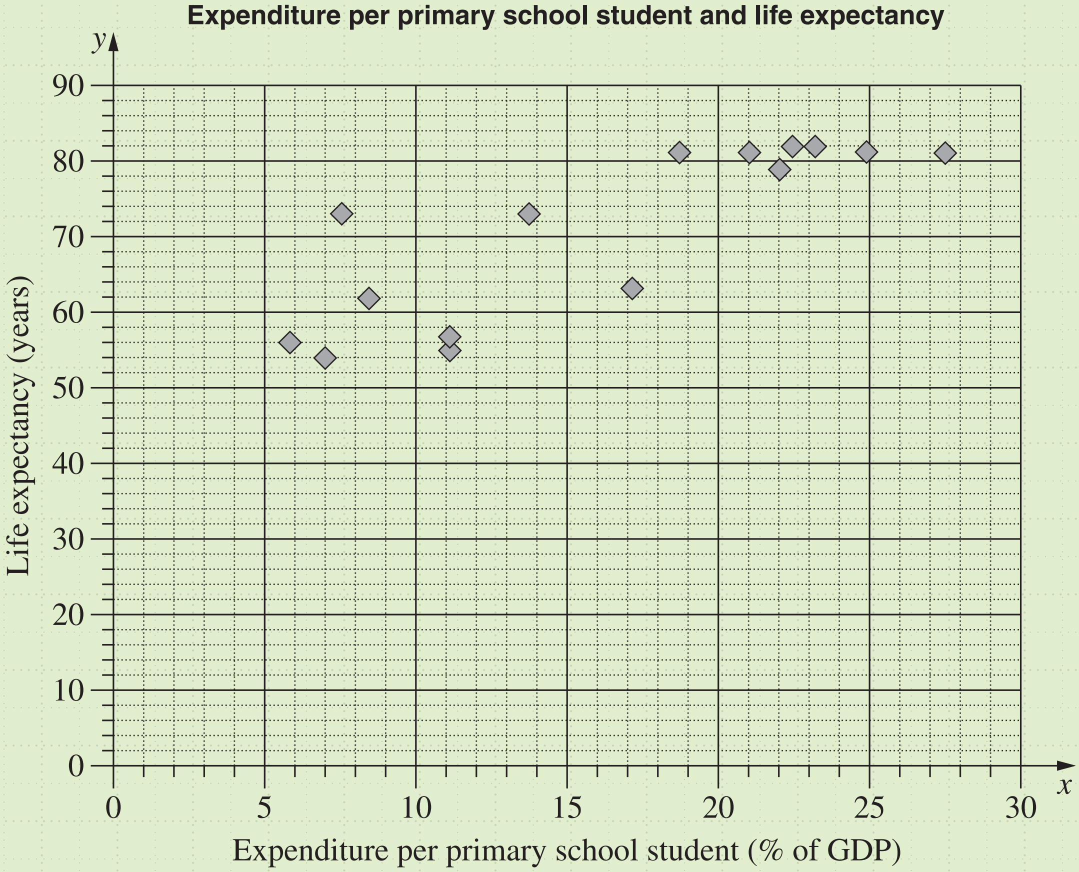
(D) increased by 1.1 hours

(E) increased by 2.3 hours

Answer: A

Example 35

[2014 General 2 HSC Q30] The scatterplot shows the relationship between expenditure per primary school student, as a percentage of a country's Gross Domestic Product (GDP), and the life expectancy in years for 15 countries.



i. For the given data, the correlation coefficient, r, is 0.83. What does this indicate about the relationship between expenditure per primary school student and life expectancy for the 15 countries?

ii. For the data representing expenditure per primary school student, Q_L is 8.4 and Q_U is 22.5.

   What is the interquartile range?

iii. Another country has an expenditure per primary school student of 47.6% of its GDP. Would this country be an outlier for this set of data? Justify your answer with calculations.

Example 35

[2014 General 2 HSC Q30] *Example 35 on the preceding page continued...*

iv. The expenditures per primary school student for the 15 countries in the 2 scatterplot are:

5.9 7 7.6 8.4 11.2 11.2 13.7 17.1 18.7
21.1 22 22.5 23.2 24.9 27.6

Complete the table below by calculating the mean, $\overline{x}$, and the standard deviation, σ_x , of these data. Calculate both values to two decimal places. The table also shows the mean, $\overline{y}$, and the standard deviation, σ_y , of life expectancy for the same 15 countries.

	Mean	Standard deviation
Expenditure per primary school student	$\overline{x} =$	$\sigma_x =$
Life expectancy	$\overline{y} = 70.73$	$\sigma_y = 10.94$

v. Using the values from the table in part (iv), show that the equation of the least-squares line of best fit is

$$y = 1.29x + 49.9$$

vi. On the scatterplot provided, draw the least-squares line of best fit,

$$y = 1.29x + 49.9$$

vii. Using this line, or otherwise, estimate the life expectancy in a country which has an expenditure per primary school student of 18% of its GDP.

viii. Why is this line NOT useful for predicting life expectancy in a country which has expenditure per primary school student of 60% of its GDP?

Answer: (ii) 14.1 (iii) Yes (iv) $\overline{x} = 16.14$, $\sigma_x = 7.03$ (vi) 73 years

Further exercises

Note: use the calculator functions and the formulas detailed in Section 3.2 and 3.3 to complete the exercises below.

(A) **Ex 9D**
- Q6-9

(A) **Ex 9E**
- Q1-2

(A) **Ex 9F**
- Q1-4

(X1) **Ex 15D**
- Q6-9

(X1) **Ex 15E**
- Q1-2

(X1) **Ex 15F**
- Q1-4

Section 4

Privacy, Bias and Ethics

Learning Goal(s)

Knowledge
Understand the effect of bias at different stages of data collection

Skills
Identify issues involving bias in real world data, particularly in the media

Understanding
Recognise the ethical issues involved with data collection and analysis

By the end of this section am I able to:

31.13 Construct, interpret and analyse scatterplots for bivariate numerical data in practical contexts

Digital data collected from a variety of sources involving digital devices, raise *privacy* and *ethical* issues.

Important note

The main *ethical issuses* that researchers must consider are:

- **Consent**: Have users agreed to the collection and use of their data?
- **Privacy**: Can the data be used to identify users?
- **Ownership**: Who owns the data and who has the right to determine what the data can be used for?
- **Data sharing and reuse**: Can the data collected be shared and can the data be used for different purposes than the purpose for which it was originally collected?

Bias can occur throughout the investigative process, from data collection through to statistical analysis and reporting.

Important note

Bias can be in favour or against an outcome, and can skew data and distort the truth of findings. This influences the validity and reliability of conclusions, which in turn impact how decisions are made.

Fill in the spaces

Bias can result from the researcher themselves, as a result of:

- selecting bias .... groups (that do not represent the target population)
- their .... (mannerisms, style of dress, speaking tone or body language)
- asking biased .... (phrase questions that influence participants' answers)
- personal .... (affects their analysis and reporting)

Example 36

How coronavirus charts can mislead us

NSW Education Standards Authority

2020 HIGHER SCHOOL CERTIFICATE EXAMINATION

Mathematics Advanced
Mathematics Extension 1
Mathematics Extension 2

REFERENCE SHEET

Measurement

Length

$$l = \frac{\theta}{360} \times 2\pi r$$

Area

$$A = \frac{\theta}{360} \times \pi r^2$$

$$A = \frac{h}{2}(a + b)$$

Surface area

$$A = 2\pi r^2 + 2\pi rh$$

$$A = 4\pi r^2$$

Volume

$$V = \frac{1}{3}Ah$$

$$V = \frac{4}{3}\pi r^3$$

Functions

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

For $ax^3 + bx^2 + cx + d = 0$:

$$\alpha + \beta + \gamma = -\frac{b}{a}$$

$$\alpha\beta + \alpha\gamma + \beta\gamma = \frac{c}{a}$$

and $\alpha\beta\gamma = -\frac{d}{a}$

Relations

$$(x - h)^2 + (y - k)^2 = r^2$$

Financial Mathematics

$$A = P(1 + r)^n$$

Sequences and series

$$T_n = a + (n - 1)d$$

$$S_n = \frac{n}{2}\left[2a + (n - 1)d\right] = \frac{n}{2}(a + l)$$

$$T_n = ar^{n-1}$$

$$S_n = \frac{a(1 - r^n)}{1 - r} = \frac{a(r^n - 1)}{r - 1},\ r \neq 1$$

$$S = \frac{a}{1 - r},\ |r| < 1$$

Logarithmic and Exponential Functions

$$\log_a a^x = x = a^{\log_a x}$$

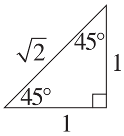
$$\log_a x = \frac{\log_b x}{\log_b a}$$

$$a^x = e^{x \ln a}$$

Trigonometric Functions

$\sin A = \dfrac{\text{opp}}{\text{hyp}}, \quad \cos A = \dfrac{\text{adj}}{\text{hyp}}, \quad \tan A = \dfrac{\text{opp}}{\text{adj}}$

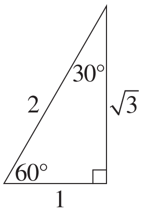
$A = \dfrac{1}{2}ab\sin C$



$\dfrac{a}{\sin A} = \dfrac{b}{\sin B} = \dfrac{c}{\sin C}$

$c^2 = a^2 + b^2 - 2ab\cos C$

$\cos C = \dfrac{a^2 + b^2 - c^2}{2ab}$



$l = r\theta$

$A = \dfrac{1}{2}r^2\theta$

Trigonometric identities

$\sec A = \dfrac{1}{\cos A}, \ \cos A \neq 0$

$\operatorname{cosec} A = \dfrac{1}{\sin A}, \ \sin A \neq 0$

$\cot A = \dfrac{\cos A}{\sin A}, \ \sin A \neq 0$

$\cos^2 x + \sin^2 x = 1$

Compound angles

$\sin(A + B) = \sin A\cos B + \cos A\sin B$

$\cos(A + B) = \cos A\cos B - \sin A\sin B$

$\tan(A + B) = \dfrac{\tan A + \tan B}{1 - \tan A\tan B}$

If $t = \tan\dfrac{A}{2}$ then $\sin A = \dfrac{2t}{1 + t^2}$

$\cos A = \dfrac{1 - t^2}{1 + t^2}$

$\tan A = \dfrac{2t}{1 - t^2}$

$\cos A\cos B = \dfrac{1}{2}\left[\cos(A - B) + \cos(A + B)\right]$

$\sin A\sin B = \dfrac{1}{2}\left[\cos(A - B) - \cos(A + B)\right]$

$\sin A\cos B = \dfrac{1}{2}\left[\sin(A + B) + \sin(A - B)\right]$

$\cos A\sin B = \dfrac{1}{2}\left[\sin(A + B) - \sin(A - B)\right]$

$\sin^2 nx = \dfrac{1}{2}(1 - \cos 2nx)$

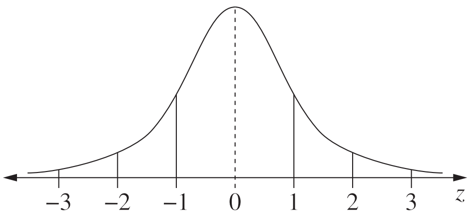
$\cos^2 nx = \dfrac{1}{2}(1 + \cos 2nx)$

Statistical Analysis

$z = \dfrac{x - \mu}{\sigma}$

An outlier is a score
less than $Q_1 - 1.5 \times IQR$
or
more than $Q_3 + 1.5 \times IQR$

Normal distribution



- approximately 68% of scores have z-scores between −1 and 1
- approximately 95% of scores have z-scores between −2 and 2
- approximately 99.7% of scores have z-scores between −3 and 3

$E(X) = \mu$

$\operatorname{Var}(X) = E\left[(X - \mu)^2\right] = E\left(X^2\right) - \mu^2$

Probability

$P(A \cap B) = P(A)P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A|B) = \dfrac{P(A \cap B)}{P(B)}, \ P(B) \neq 0$

Continuous random variables

$P(X \le x) = \displaystyle\int_a^x f(x)\,dx$

$P(a < X < b) = \displaystyle\int_a^b f(x)\,dx$

Binomial distribution

$P(X = r) = {}^nC_r p^r(1 - p)^{n-r}$

$X \sim \operatorname{Bin}(n, p)$

$\Rightarrow \ P(X = x)$

$= \dbinom{n}{x} p^x(1 - p)^{n-x}, \ x = 0, 1, \ldots, n$

$E(X) = np$

$\operatorname{Var}(X) = np(1 - p)$

Differential Calculus

Function	Derivative
$y = f(x)^n$	$\dfrac{dy}{dx} = nf'(x)\left[f(x)\right]^{n-1}$
$y = uv$	$\dfrac{dy}{dx} = u\dfrac{dv}{dx} + v\dfrac{du}{dx}$
$y = g(u)$ where $u = f(x)$	$\dfrac{dy}{dx} = \dfrac{dy}{du} \times \dfrac{du}{dx}$
$y = \dfrac{u}{v}$	$\dfrac{dy}{dx} = \dfrac{v\dfrac{du}{dx} - u\dfrac{dv}{dx}}{v^2}$
$y = \sin f(x)$	$\dfrac{dy}{dx} = f'(x)\cos f(x)$
$y = \cos f(x)$	$\dfrac{dy}{dx} = -f'(x)\sin f(x)$
$y = \tan f(x)$	$\dfrac{dy}{dx} = f'(x)\sec^2 f(x)$
$y = e^{f(x)}$	$\dfrac{dy}{dx} = f'(x)e^{f(x)}$
$y = \ln f(x)$	$\dfrac{dy}{dx} = \dfrac{f'(x)}{f(x)}$
$y = a^{f(x)}$	$\dfrac{dy}{dx} = (\ln a)f'(x)a^{f(x)}$
$y = \log_a f(x)$	$\dfrac{dy}{dx} = \dfrac{f'(x)}{(\ln a)f(x)}$
$y = \sin^{-1} f(x)$	$\dfrac{dy}{dx} = \dfrac{f'(x)}{\sqrt{1-\left[f(x)\right]^2}}$
$y = \cos^{-1} f(x)$	$\dfrac{dy}{dx} = -\dfrac{f'(x)}{\sqrt{1-\left[f(x)\right]^2}}$
$y = \tan^{-1} f(x)$	$\dfrac{dy}{dx} = \dfrac{f'(x)}{1+\left[f(x)\right]^2}$

Integral Calculus

$$\int f'(x)\left[f(x)\right]^n dx = \frac{1}{n+1}\left[f(x)\right]^{n+1} + c$$

where $n \neq -1$

$$\int f'(x)\sin f(x)\,dx = -\cos f(x) + c$$

$$\int f'(x)\cos f(x)\,dx = \sin f(x) + c$$

$$\int f'(x)\sec^2 f(x)\,dx = \tan f(x) + c$$

$$\int f'(x)e^{f(x)}\,dx = e^{f(x)} + c$$

$$\int \frac{f'(x)}{f(x)}\,dx = \ln\left|f(x)\right| + c$$

$$\int f'(x)a^{f(x)}\,dx = \frac{a^{f(x)}}{\ln a} + c$$

$$\int \frac{f'(x)}{\sqrt{a^2-\left[f(x)\right]^2}}\,dx = \sin^{-1}\frac{f(x)}{a} + c$$

$$\int \frac{f'(x)}{a^2+\left[f(x)\right]^2}\,dx = \frac{1}{a}\tan^{-1}\frac{f(x)}{a} + c$$

$$\int u\frac{dv}{dx}\,dx = uv - \int v\frac{du}{dx}\,dx$$

$$\int_a^b f(x)\,dx$$

$$\approx \frac{b-a}{2n}\left\{f(a) + f(b) + 2\left[f\left(x_1\right) + \cdots + f\left(x_{n-1}\right)\right]\right\}$$

where $a = x_0$ and $b = x_n$

Combinatorics

$${}^{n}P_{r} = \frac{n!}{(n-r)!}$$

$$\binom{n}{r} = {}^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

$$(x+a)^{n} = x^{n} + \binom{n}{1}x^{n-1}a + \cdots + \binom{n}{r}x^{n-r}a^{r} + \cdots + a^{n}$$

Vectors

$$\left|\underset{\sim}{u}\right| = \left|x\underset{\sim}{i} + y\underset{\sim}{j}\right| = \sqrt{x^{2}+y^{2}}$$

$$\underset{\sim}{u}\cdot\underset{\sim}{v} = \left|\underset{\sim}{u}\right|\left|\underset{\sim}{v}\right|\cos\theta = x_1x_2 + y_1y_2,$$

where $\underset{\sim}{u} = x_1\underset{\sim}{i} + y_1\underset{\sim}{j}$

and $\underset{\sim}{v} = x_2\underset{\sim}{i} + y_2\underset{\sim}{j}$

$$\underset{\sim}{r} = \underset{\sim}{a} + \lambda\underset{\sim}{b}$$

Complex Numbers

$$z = a + ib = r(\cos\theta + i\sin\theta) = re^{i\theta}$$

$$\left[r(\cos\theta + i\sin\theta)\right]^{n} = r^{n}(\cos n\theta + i\sin n\theta) = r^{n}e^{in\theta}$$

Mechanics

$$\frac{d^{2}x}{dt^{2}} = \frac{dv}{dt} = v\frac{dv}{dx} = \frac{d}{dx}\left(\frac{1}{2}v^{2}\right)$$

$$x = a\cos(nt+\alpha) + c$$

$$x = a\sin(nt+\alpha) + c$$

$$\ddot{x} = -n^{2}(x-c)$$