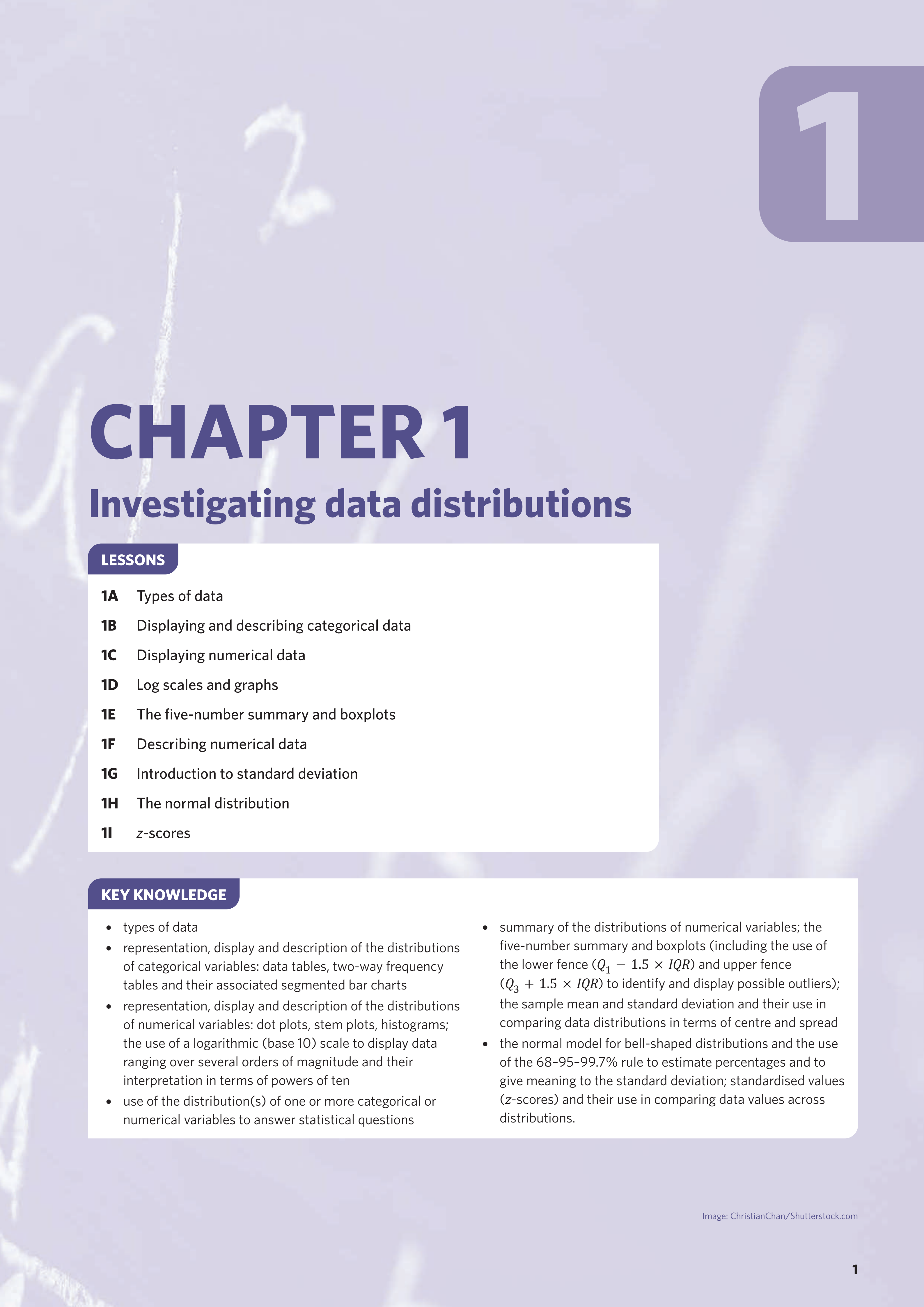


1

CHAPTER 1

Investigating data distributions

LESSONS

1A Types of data

1B Displaying and describing categorical data

1C Displaying numerical data

1D Log scales and graphs

1E The five-number summary and boxplots

1F Describing numerical data

1G Introduction to standard deviation

1H The normal distribution

1I *z*-scores

KEY KNOWLEDGE

- types of data
- representation, display and description of the distributions of categorical variables: data tables, two-way frequency tables and their associated segmented bar charts
- representation, display and description of the distributions of numerical variables: dot plots, stem plots, histograms; the use of a logarithmic (base 10) scale to display data ranging over several orders of magnitude and their interpretation in terms of powers of ten
- use of the distribution(s) of one or more categorical or numerical variables to answer statistical questions
- summary of the distributions of numerical variables; the five-number summary and boxplots (including the use of the lower fence ($Q_1 - 1.5 \times IQR$) and upper fence ($Q_3 + 1.5 \times IQR$) to identify and display possible outliers); the sample mean and standard deviation and their use in comparing data distributions in terms of centre and spread
- the normal model for bell-shaped distributions and the use of the 68–95–99.7% rule to estimate percentages and to give meaning to the standard deviation; standardised values (*z*-scores) and their use in comparing data values across distributions.

Image: ChristianChan/Shutterstock.com

1A Types of data

STUDY DESIGN DOT POINT

- types of data

1A 1B 1C 1D 1E 1F 1G 1H 1I

KEY SKILLS

During this lesson, you will be:

- classifying data as categorical or numerical
- classifying categorical data as nominal or ordinal
- classifying numerical data as discrete or continuous.

KEY TERMS

- Data
- Categorical data
- Numerical data
- Nominal data
- Ordinal data
- Discrete data
- Continuous data

In the Information Age, data is becoming increasingly more important to everyday life. Classifying data into data types is necessary before analysis can be performed, or the most appropriate data visualisations can be constructed.

Classifying data as categorical or numerical

Data is a set of values, words or responses, that is collected and ordered by variables.

Data that can be organised into categories or groups is known as **categorical data**. It is also referred to as qualitative data, as it represents a quality or attribute.

Data that can be counted or measured is known as **numerical data**. It is also referred to as quantitative data, as it represents a quantity.

Worked example 1

Classify the following variables as either categorical or numerical.

a. *type of pasta*

Explanation

The variable *type of pasta* is categorised into different pasta types such as gnocchi, fettuccine, spaghetti or lasagne.

Answer

Categorical

b. *number of candles*

Explanation

The variable *number of candles* is counted.

Answer

Numerical

Classifying categorical data as nominal or ordinal

Categorical data can be further classified as either nominal or ordinal.

Categorical data that cannot be sorted into a logical ordered list or hierarchy is called **nominal data**. For example, *type of bread* (white bread, multigrain, sourdough) has no inherent ranking system and is classified as nominal categorical data.

Categorical data that can be ordered into a logical ordered list or hierarchy is called **ordinal data**. For example, *drink size* (small, medium, large) can be ordered such that medium is greater than small, and large is greater than medium. This is an inherent ranking system, so it is classified as ordinal categorical data.

Worked example 2

Classify the following categorical variables as either nominal or ordinal.

a. *type of shoe* (runners, boots, sandals, slides)

Explanation

The categories within the variable *type of shoe* cannot be inherently ordered.

Answer

Nominal

b. *shirt size* (small, medium, large)

Explanation

The categories within the variable *shirt size* can be inherently ordered (small to medium to large).

Answer

Ordinal

Classifying numerical data as discrete or continuous

Numerical variables can be further classified as either discrete or continuous.

Numerical data that can only consist of a set of fixed values within a range is called **discrete data**. Discrete data usually consists of whole numbers and would typically be collected by counting. For example, the *number of steps* taken in a day can only be represented by whole numbers starting from zero, and is classified as discrete numerical data.

Numerical data that can consist of any value within a range is called **continuous data**. Continuous data usually consists of both whole numbers and decimals and would typically be collected by measuring. For example, the *distance* (km) walked in a day is classified as continuous numerical data as it is measured and can consist of any positive value, such as 5.1, 5.01 or even 5.001. Continuous data that has been rounded to the nearest whole number is still considered to be continuous.

Worked example 3

Classify the following numerical variables as either discrete or continuous.

a. *length* (m)

Explanation

The variable *length* (m) can be expressed in decimals and can consist of any value measured on a continuous scale.

Answer

Continuous

b. *number of tennis racquets*

Explanation

The variable *number of tennis racquets* cannot be expressed in decimals and can only be counted.

Answer

Discrete

Exam question breakdown

VCAA 2016 Exam 1 Data analysis Q2

The variables *blood pressure* (low, normal, high) and *age* (under 50 years, 50 years or over) are

A. both nominal variables.

B. both ordinal variables.

C. a nominal variable and an ordinal variable respectively.

D. an ordinal variable and a nominal variable respectively.

E. a continuous variable and an ordinal variable respectively.

Explanation

Step 1: Classify the variable *blood pressure* (low, normal, high).

The variable *blood pressure* has three categories, low, medium and high. As such, this is a categorical variable.

These categories can be sorted into ascending or descending order. Therefore, *blood pressure* (low, normal, high) can be further classified as an ordinal variable.

Step 2: Classify the variable *age* (under 50 years, 50 years or over).

The variable *age* has two categories; 'under 50 years' and '50 years or over'. As such, this is a categorical variable.

These categories can also be sorted into ascending or descending order. Therefore, *age* (under 50 years, 50 years or over) can be further classified as an ordinal variable.

Answer

B

31% of students answered this question correctly.

45% of students incorrectly chose option D, as they identified the variable *age* (under 50 years, 50 years or over) as a nominal variable. The variable *age* is ordinal since one group of people can be classified as younger than the other group, creating an inherent order between the two categories.

1A Questions

Classifying data as categorical or numerical

1. Which of the following variables is categorical?

A. *number of lamps*

B. *number of wardrobes*

C. *cost of a house*

D. *type of kitchen*

2. Which of the following variables is numerical?

A. *number of teachers*

B. *type of cake*

C. *type of painting*

D. *laptop brand* (1 = Apple, 2 = ASUS, 3 = HP, 4 = other)

3. Classify the following variables as either categorical or numerical.

a. *age*

b. *exam difficulty* (1 = easy, 2 = medium, 3 = hard)

Classifying categorical data as nominal or ordinal

4. Which of the following categorical variables is nominal?

A. *clay quality* (low, medium, high)

B. *class participation* (low, moderate, high)

C. *weather forecast* (sunny, clear, cloudy, raining)

D. *level of processing* (shallow, moderate, deep)

5. Which of the following categorical variables is ordinal?

A. *keyboard switch type* (blue, red, brown)

B. *difficulty ranking* (1 = easy, 2 = moderate, 3 = hard)

C. *personality type* (INTP, ISTJ, ENTJ, etc...)

D. *favourite ice cream flavour* (black sesame, green tea, vanilla)

6. Classify the following categorical variables as either nominal or ordinal.

a. *type of car* (1 = sedan, 2 = sports, 3 = convertible, 4 = other)

b. *assessment grade* (A, B, C, D, E, F)

Classifying numerical data as discrete or continuous

7. Which of the following numerical variables is discrete?

A. *time elapsed*

B. *height*

C. *number of keyboards*

D. *volume of* CO_2 *output*

8. Which of the following numerical variables is continuous?

A. *student enrolments*

B. *tennis tournaments won*

C. *number of dogs*

D. *bone mass*

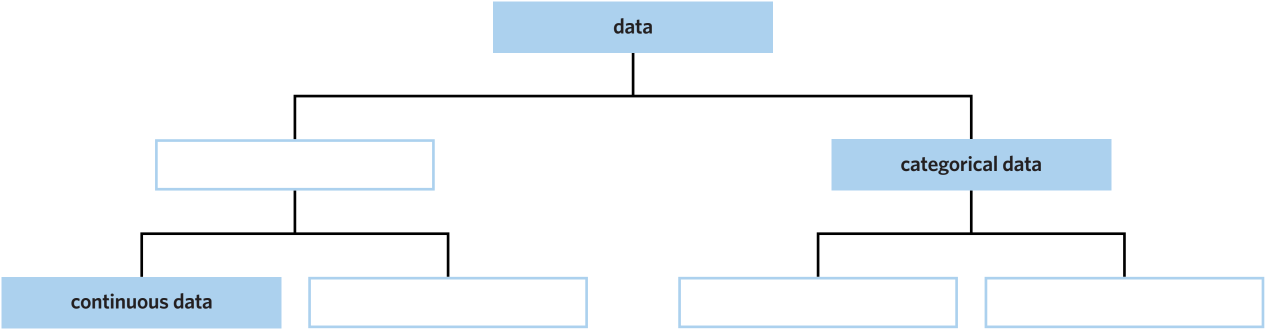
9. Classify the following numerical variables as either discrete or continuous.

a. The *number of parrots* found in different rainforests.

b. The *haemoglobin count* of a group of people, in (g/dl).

Joining it all together

10. Fill in the gaps with the following terms: nominal data, discrete data, numerical data, and ordinal data.



11. Classify the following variables as either nominal, ordinal, discrete or continuous.

a. *car brand* (1 = Toyota, 2 = Holden, 3 = Ford, 4 = other)

b. *number of employees*

c. *weight of textbook* (kg)

d. *height of basketball players* (cm)

e. *perfume brand*

f. *exam grades* (HD = high distinction, D = distinction, C = credit, P = pass, N = fail)

g. *student number*

h. *number of users*

i. *postcode*

j. *user rating* (1 = not satisfactory, 2 = neutral, 3 = satisfactory)

12. A tennis coach collected data on the *number of tennis racquets used* and the *serve speed* (km/h) for several tennis players for the upcoming Australian Open.

a. Which of the two variables is continuous?

b. Which of the two variables is discrete?

13. A software company wants to see if they need to upgrade their program. They conduct a survey where the participants are asked to comment on the statement 'The program is easy to navigate'. They collect the responses under the variable *response* (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree).

The program is easy to navigate

	1	2	3	4	5	
strongly disagree	○	○	○	○	○	strongly agree

What type of data are they collecting?

A. Nominal **B.** Ordinal **C.** Discrete **D.** Continuous

Exam practice

14. The table shows the *day number* and the *minimum temperature*, in degrees Celsius, for 15 consecutive days in May 2017.

day number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
***minimum temperature* (°C)**	12.7	11.8	10.7	9.0	6.0	7.0	4.1	4.8	9.2	6.7	7.5	8.0	8.6	9.8	7.7

Which of the two variables in this data set is an ordinal variable? (1 MARK)

VCAA 2019 Exam 2 Data analysis Q1a

81% of students answered this question correctly.

15. Data relating to the following five variables was collected from insects that were caught overnight in a trap:

- *colour*
- *name of species*
- *number of wings*
- *body length* (in millimetres)
- *body weight* (in milligrams)

The number of these variables that are discrete variables is

A. 1 **B.** 2 **C.** 3
D. 4 **E.** 5

VCAA 2020 Exam 1 Data analysis Q7

69% of students answered this question correctly.

16. In the sport of heptathlon, athletes compete in seven events.

These events are the 100 m hurdles, high jump, shot-put, javelin, 200 m run, 800 m run and long jump.

Fifteen female athletes competed to qualify for the heptathlon at the Olympic Games.

Their results for three of the heptathlon events – high jump, shot-put and javelin – are shown in the table.

athlete number	***high jump* (metres)**	***shot-put* (metres)**	***javelin* (metres)**
1	1.76	15.34	41.22
2	1.79	16.96	42.41
3	1.83	13.87	46.53
4	1.82	14.23	40.53
5	1.87	13.78	40.62
6	1.73	14.50	45.62
7	1.68	15.08	42.33
8	1.82	13.13	40.88
9	1.83	14.22	39.22
10	1.87	13.62	42.51
11	1.87	12.01	42.75
12	1.80	12.88	38.12
13	1.83	12.68	42.65
14	1.87	12.45	41.32
15	1.78	11.31	42.88

Write down the number of numerical variables in the table. (1 MARK)

VCAA 2021 Exam 2 Data analysis Q1a

52% of students answered this question correctly.

17. The variables *number of moths* (less than 250, 250–500, more than 500) and *trap type* (sugar, scent, light) are

A. both nominal variables.

B. both ordinal variables.

C. a numerical variable and a categorical variable respectively.

D. a nominal variable and an ordinal variable respectively.

E. an ordinal variable and a nominal variable respectively.

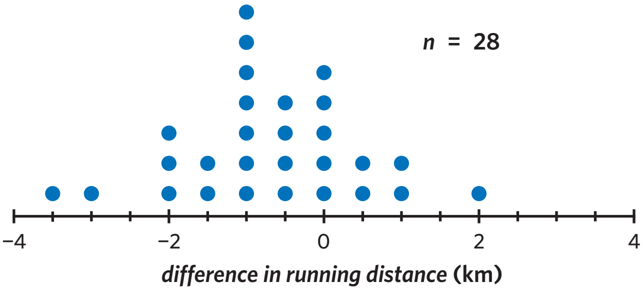
VCAA 2017 Exam 1 Data analysis Q7

46% of students answered this question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

18. Ashleigh and Savannah are training to run a marathon by running as far as they can inside 3 hours and 30 minutes. The dot plot displays the *difference in distance* run by Ashleigh in relation to Savannah (i.e. 0.5 means Ashleigh ran 500 m more than Savannah, while −0.5 means Ashleigh ran 500 m less than Savannah). They ran together 28 times.



The percentage of days in which Ashleigh ran one less kilometre than Savannah is

A. 7.1% **B.** 10.7% **C.** 14.3% **D.** 25.0% **E.** 28.0%

Adapted from VCAA 2018 Exam 1 Data analysis Q1

Recursion and financial modelling *Year 11 content*

19. Arthur gets $1000 for his birthday and wants to save his money. He opens a savings account and deposits his $1000. The account earns interest at a rate of 3% per annum, compounding annually.

Let V_n be the value of Arthur's account n years after he initially deposits his money.

The expected growth of Athur's savings account can be modelled by

A. $V_0 = 1000, \quad V_{n+1} = V_n + 30$

B. $V_1 = 1030, \quad V_{n+1} = V_n + 30$

C. $V_0 = 1000, \quad V_{n+1} = 1.03\, V_n$

D. $V_0 = 1030, \quad V_{n+1} = 1.03\, V_n$

E. $V_1 = 1000, \quad V_{n+1} = 1.3\, V_n$

Adapted from VCAA 2015 Exam 1 Number patterns Q3

Data analysis *Year 11 content*

20. The number of cars that park in a particular car park on each day of one week are counted and recorded in the following table.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
number of cars	103	84	92	79	93	64	48

From the information given, determine

a. the range. (1 MARK)

b. the percentage of days that had less than 90 cars parked, correct to one decimal place. (1 MARK)

Adapted from VCAA 2017 Exam 2 Data analysis Q1

1B Displaying and describing categorical data

STUDY DESIGN DOT POINTS

- representation, display and description of the distributions of categorical variables: data tables, two-way frequency tables and their associated segmented bar charts
- use of the distribution(s) of one or more categorical or numerical variables to answer statistical questions

1A 1B 1C 1D 1E 1F 1G 1H 1I

KEY SKILLS

During this lesson, you will be:

- constructing frequency tables
- constructing bar charts
- constructing segmented bar charts
- describing the distribution of categorical data.

KEY TERMS

- Frequency table
- Percentage frequency
- Bar chart
- Segmented bar chart
- Percentage segmented bar chart
- Mode

Lists of categorical information can be converted into tables, graphs and charts so that they can be easily read and interpreted. These displays can be used to identify the number, or percentage, of data for each category, as well as the most frequently occurring category.

Constructing frequency tables

A **frequency table** is a table that tallies how often each value in a data set occurs. This is the first step in making a set of data easier to summarise and analyse.

Data can be recorded within a frequency table as either frequency or **percentage frequency**. The percentage frequency is the proportion of times each value or category occurs in relation to the entire data set, represented as a percentage.

$$\textit{percentage frequency} = \frac{\textit{frequency}}{\textit{total frequency}} \times 100$$

Worked example 1

The students in a prep class were asked the question, ‘Would you describe your teacher’s height as short, average or tall?’. Their responses were as follows:

average	short	tall	tall	short	short
average	average	tall	short	average	short
average	average	average	tall	tall	short
average	tall	tall	average	average	tall

Use this data to create a frequency table displaying both frequency and percentage frequency, correct to the nearest decimal place.

Continues →

Explanation

Step 1: Set up a frequency table.

The table should have 3 columns for the variable collected, and the frequency as a number and percentage. There should be an appropriate number of rows to include all the categories. Finally, a row should be included for the total.

teacher's height	**frequency**	
	number	**%**
short		
average		
tall		
total		

Step 2: Fill in the frequency number column by counting from the data set, including the total.

teacher's height	**frequency**	
	number	**%**
short	6	
average	10	
tall	8	
total	24	

Step 3: Calculate the frequency as a percentage for each category, making sure the percentages add up to 100.

Note: When percentages have been rounded, they may not add up to exactly 100. In these situations this is okay, as long as the rounding has been done accurately.

Remember that the question asks for percentages given to the nearest decimal place.

$$\textit{percentage frequency} = \frac{\textit{frequency}}{\textit{total frequency}} \times 100$$

teacher's height	**frequency**	
	number	**%**
short	6	$\frac{6}{24} \times 100 = 25.0$
average	10	$\frac{10}{24} \times 100 \approx 41.7$
tall	8	$\frac{8}{24} \times 100 \approx 33.3$
total	24	100.0

Answer

teacher's height	**frequency**	
	number	**%**
short	6	25.0
average	10	41.7
tall	8	33.3
total	24	100.0

Constructing bar charts

A **bar chart** is a graphical display that is commonly used to display categorical data. The frequency or percentage frequency of each category is represented by columns of varied height. Spaces are included between columns to indicate that the categories are separate.

Worked example 2

24 students in a prep class were asked the question, ‘Would you describe your *teacher’s height* as short, average or tall?’. Their responses are recorded in the frequency table shown.

Use the frequency table to construct a frequency bar chart.

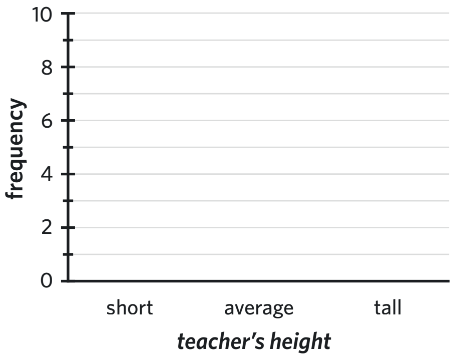
teacher’s height	**frequency**	
	number	**%**
short	6	25.0
average	10	41.7
tall	8	33.3
total	24	100.0

Explanation

Step 1: Construct axes with the ‘frequency’ on the vertical axis and ‘*teacher’s height*’ on the horizontal axis.

The vertical axis should at least extend to the maximum value.

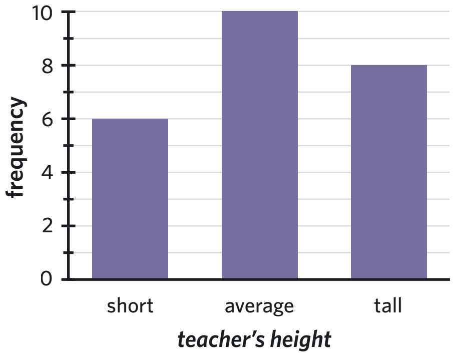
The horizontal axis should include labels for each of the categories.



Step 2: Draw vertical columns for each category according to their value in the frequency table.

Remember that each column should be separated by a gap.

Answer



Constructing segmented bar charts

A **segmented bar chart** is a variation of a bar chart with each category stacked into one column. They are particularly useful for comparing the distribution of categories across different sets of data. This will be explored further later.

Each category within a segmented bar chart has its own segment, with no gaps between segments. The height of each segment indicates the frequency of each category. A legend indicates which segments of the bar relate to which categories. Segmented bar charts can also be constructed for the percentage frequency of a data set. This is called a **percentage segmented bar chart**.

Worked example 3

24 students in a prep class were asked the question, ‘Would you describe your *teacher’s height* as short, average or tall?’. Their responses are recorded in the frequency table shown.

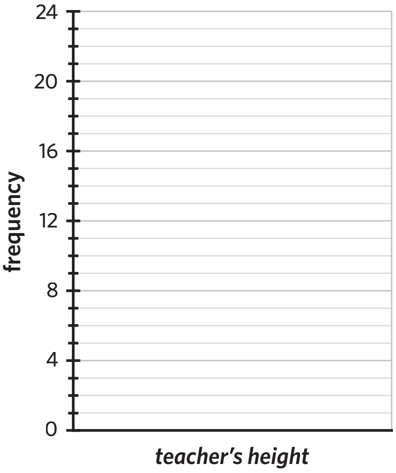
teacher’s height	**frequency**	
	number	**%**
short	6	25.0
average	10	41.7
tall	8	33.3
total	24	100.0

a. Use the frequency table to construct a segmented bar chart.

Explanation

Step 1: Construct axes with the ‘frequency’ on the vertical axis and ‘*teacher’s height*’ on the horizontal axis.

The vertical axis should at least extend to the total frequency.



Step 2: Construct the column by adding the value of each segment.

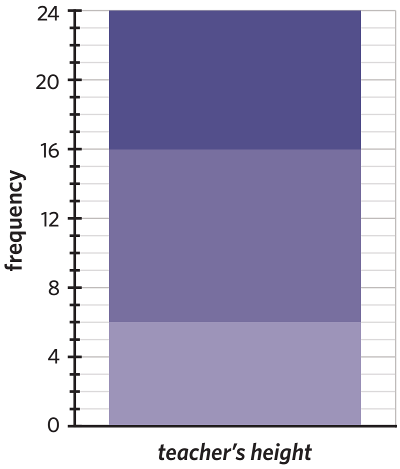
For this segmented bar chart, we will go from short,to average, to tall.

The ‘short’ segment should end at 6.

The ‘average’ segment should end at $6 + 10 = 16$.

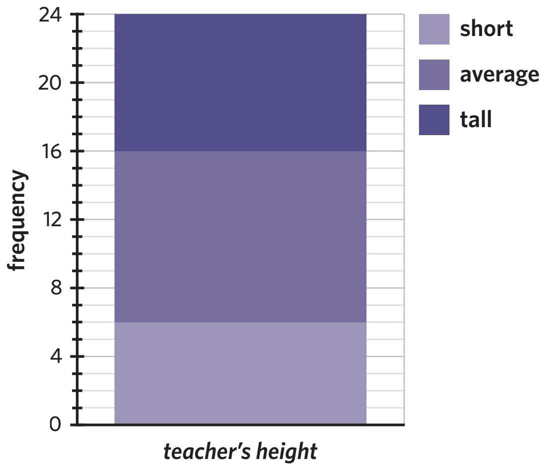
The ‘tall’ segment should end at $16 + 8 = 24$.

Ensure each segment is clearly defined.



Step 3: Add a legend so the graph can be interpreted correctly.

Answer



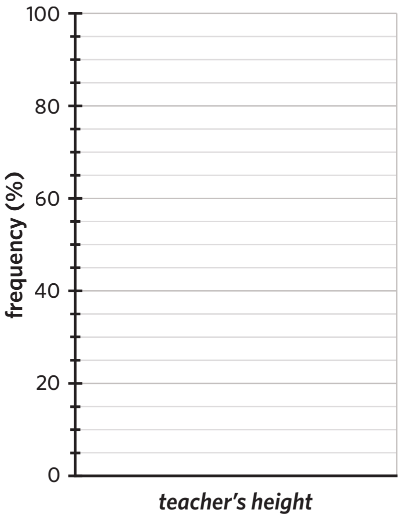
Continues →

b. Use the frequency table to construct a percentage segmented bar chart.

Explanation

Step 1: Construct axes with the frequency as a percentage on the vertical axis and *'teacher's height*' on the horizontal axis.

The vertical axis should extend to 100%.



Step 2: Construct the column by adding the percentage of each segment.

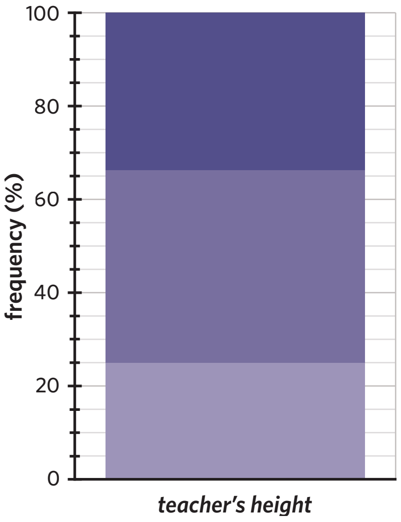
For this percentage segmented bar chart, we will go from short, to average, to tall.

The 'short' segment should end at 25.

The 'average' segment should end at 25 + 41.7 = 66.7.

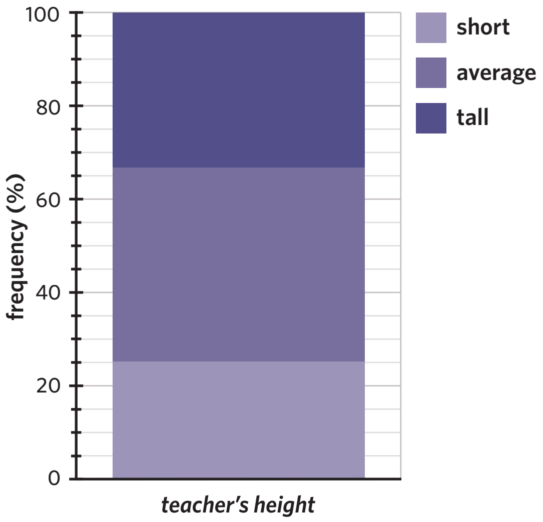
The 'tall' segment should end at 66.7 + 33.3 = 100.

Ensure each segment is clearly defined.



Step 3: Add a legend so the graph can be interpreted correctly.

Answer



Describing the distribution of categorical data

When describing data, the mean, median and mode are often mentioned as measures of centre. This is the middle, or 'average' value of a distribution. The mode is the only available measure of centre for categorical data as the mean and median only apply to numerical data. The **mode** is the most frequently occurring value in the data set. It can be identified from a bar chart or segmented bar chart by looking at the column or segment with the greatest vertical height.

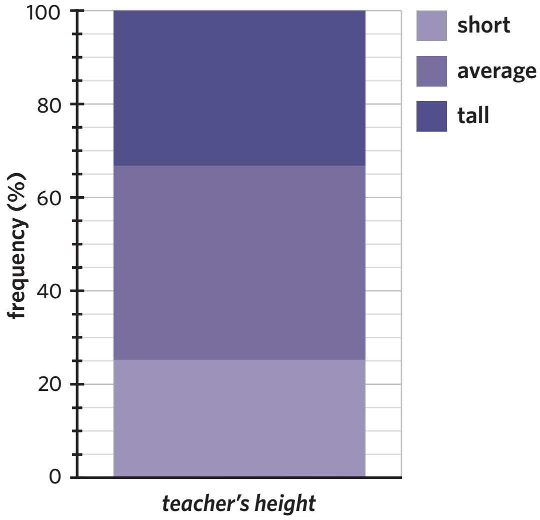
An interpretation of frequency tables, bar charts and segmented bar charts often involves writing a report which can:

- summarise the data type and the number of values represented in the data set
- identify the modal category (if it is obvious)
- compare the percentage frequencies of different categories.

In larger data sets, not all categories need to be mentioned. It might be easier to draw attention to the largest and smallest columns.

Worked example 4

24 students in a prep class were asked the question, 'Would you describe your *teacher's height* as short, average or tall?'. Their responses are shown in the given percentage segmented bar chart.



a. Find the modal category of the data set.

Explanation

Identify the segment with the greatest vertical height.

Answer

Average

b. Describe the distribution of the data set.

Explanation

Consider the elements to be included in the report describing the distribution.

- Number of people surveyed
- Modal category
- Other significant percentages

Answer

24 prep students were surveyed on how tall they thought their teacher was. The most common response was average, accounting for 41.7% of responses, while 25% said their teacher was short, and 33.3% said their teacher was tall.

1B Questions

Constructing frequency tables

1. A group of people were asked whether they preferred coffee, tea, or neither in the morning. Their results are displayed in the following frequency table, with percentages rounded to the nearest whole number.

drink preference	**frequency**	
	number	**%**
coffee	19	59
tea	5	16
neither	8	25
total	32	100

Which of the following statements is true?

A. 19 people were surveyed, and 59% preferred coffee.

B. 32 people were surveyed, and 8% preferred neither.

C. 32 people were surveyed, and 59% preferred coffee.

D. 32 people were surveyed, and 16 of them preferred tea.

2. 20 members of the Italian Club were asked what their *favourite type of pasta* is. Their results were as follows:

penne	penne	spaghetti	fettuccine	penne
fettuccine	macaroni	spaghetti	penne	spaghetti
spaghetti	fettuccine	penne	macaroni	penne
spaghetti	fettuccine	spaghetti	penne	fettuccine

Use these results to construct a frequency table including frequencies and percentages.

Constructing bar charts

3. A class of 312 Year 12 boys were asked their *shoe size*. Their results are recorded in the given bar chart.

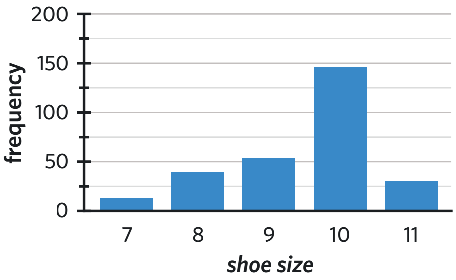
The number of Year 12 boys with a size 9 shoe is closest to

A. 10

B. 30

C. 45

D. 55



4. The *shirt sizes* (extra small, small, medium, large, extra large) of 49 people are displayed in the given frequency table. Percentages are rounded to the nearest decimal place.

a. Use the frequency table to construct a frequency bar chart.

b. Use the frequency table to construct a percentage frequency bar chart.

shirt size	**frequency**	
	number	**%**
extra small	3	6.1
small	14	28.6
medium	17	34.7
large	9	18.4
extra large	6	12.2
total	49	100.0

Constructing segmented bar charts

5. A group of people were asked what their preferred *streaming service* was. Their responses are shown in the frequency segmented bar chart shown

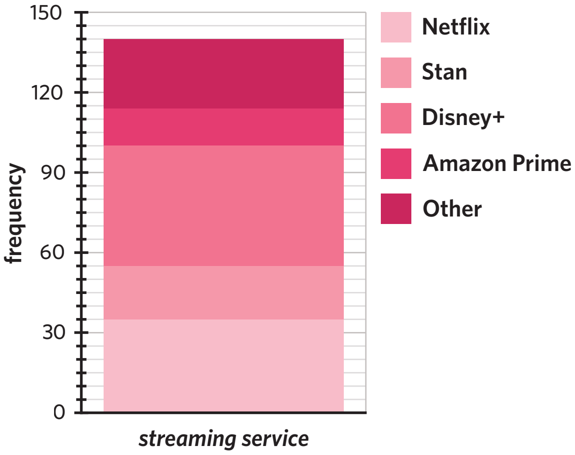
Which of the following statements is false?

A. 20 people prefer Stan.

B. More people prefer Disney+ than Netflix and Stan combined.

C. 140 people were surveyed.

D. 15 people prefer Amazon Prime.



6. 139 people were asked what their *favourite animal* was. The results are shown in the frequency table shown. Percentages have been rounded to the nearest whole number.

a. Use the data from the frequency table to construct a frequency segmented bar chart.

b. Use the data from the frequency table to construct a percentage segmented bar chart.

favourite animal	**frequency**	
	number	**%**
dog	47	34
cat	52	37
guinea pig	22	16
horse	14	10
snake	4	3
total	139	100

Describing the distribution of categorical data

7. 88 people were asked who their *favourite Avenger* was. The results are shown in the bar chart provided.

The modal superhero is

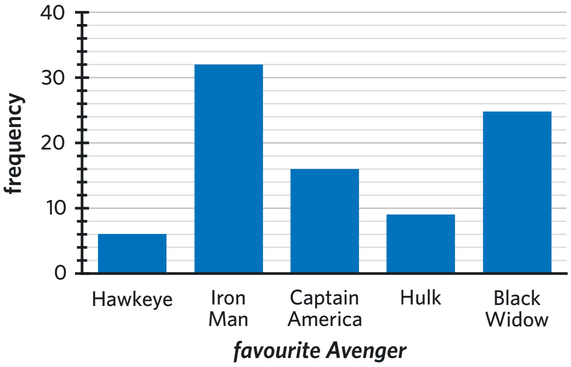
A. Hawkeye

B. Iron Man

C. Captain America

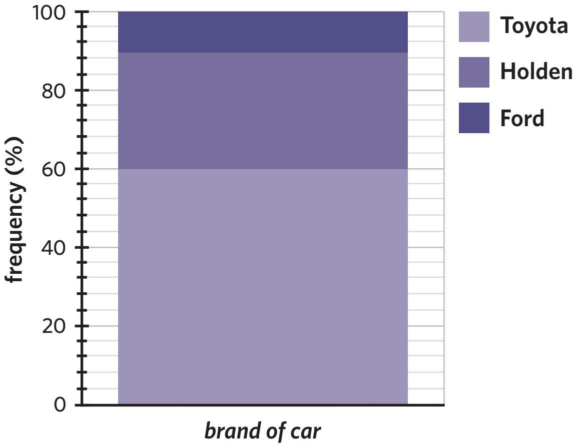
D. Hulk

E. Black Widow



8. The brands of 50 cars entering a carpark were recorded, with the results shown in the percentage segmented bar chart provided. Use this data to fill out the following report template.

The brands of __________ cars were recorded as they entered a car park. All the cars were either 'Holden', 'Ford', or 'Toyota'. The most commonly occurring brand of car was __________, accounting for __________% of all cars. The next most commonly occurring brand was __________, representing __________%. Finally, the last __________% of the cars were __________.



Joining it all together

9. A barista collected information on the type of milk that customers ordered with their coffee. The results are shown in the given percentage segmented bar chart.

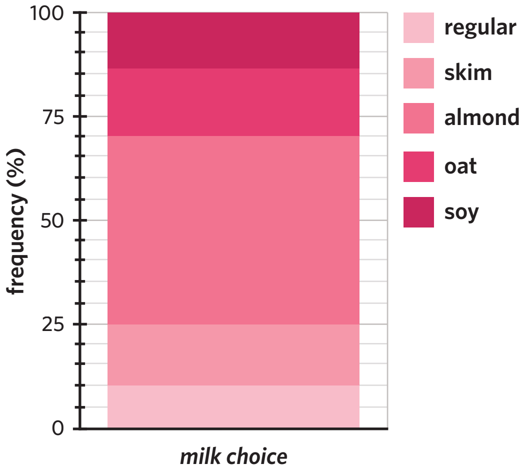
The barista remembered that 36 people ordered almond milk. The number of people that ordered regular milk is closest to

A. 8

B. 10

C. 15

D. 25



10. A group of musicians were asked who their *favourite jazz drummer* was. Their responses were as follows:

Tony Williams	Elvin Jones	Elvin Jones	Brian Blade
Tony Williams	Brian Blade	Buddy Rich	Art Blakey
Elvin Jones	Elvin Jones	Tony Williams	Buddy Rich
Buddy Rich	Art Blakey	Elvin Jones	Buddy Rich
Brian Blade	Elvin Jones	Buddy Rich	Tony Williams

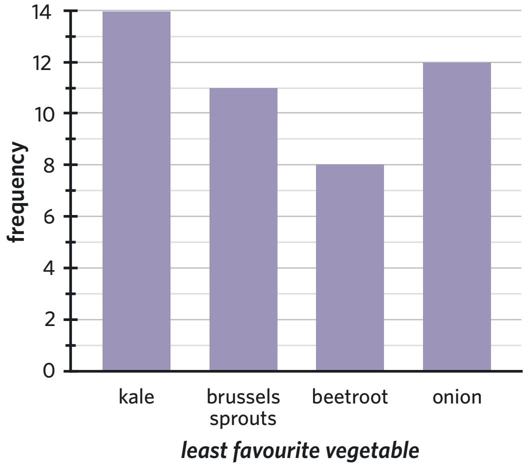
a. Use these results to construct a frequency table.

b. Using the frequency table from part **a**, construct a percentage bar chart to show the results.

c. Using the frequency table from part **a**, construct a frequency segmented bar chart to show the results.

d. Use the data to write a paragraph on the distribution of favourite jazz drummers amongst the musicians.

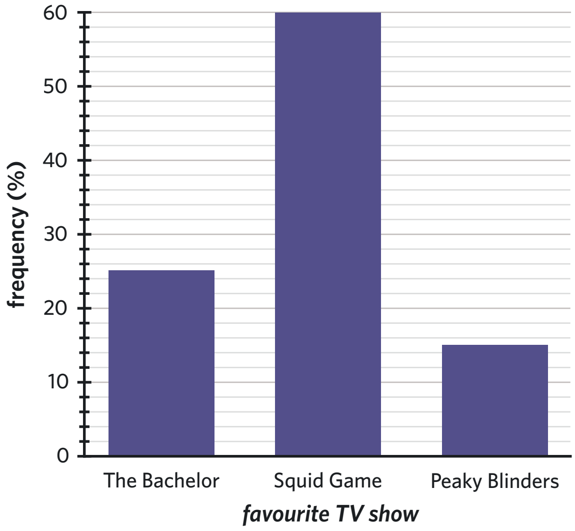
11. A group of toddlers were asked about their *least favourite vegetable*. The results are represented in the given bar chart.



Draw a percentage segmented bar chart to represent this data, correct to the nearest percentage.

12. A group of office workers were asked what their *favourite TV show* was. The results are displayed in the bar chart shown.

6 people said Peaky Blinders was their favourite show. Use this information and the bar chart to construct a frequency table that represents this information.



Exam practice

13. A study was conducted that investigated the *number of moths* caught in a sugar moth trap (less than 250, 250–500, more than 500). The results are summarised in the percentage segmented bar chart shown.

There were 300 sugar traps.

The number of sugar traps that caught less than 250 moths is closest to

A. 30

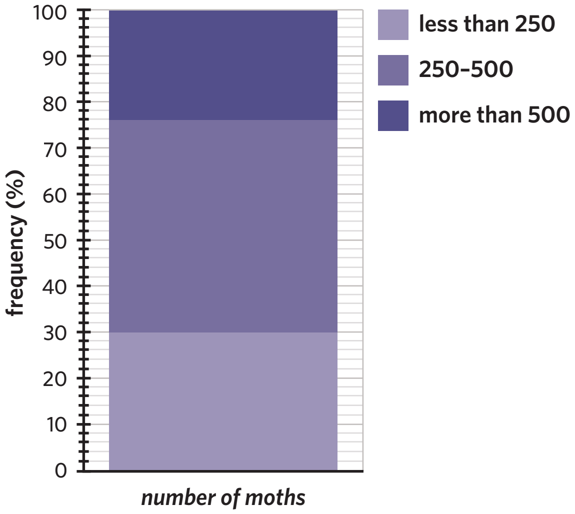
B. 90

C. 250

D. 300

E. 500

Adapted from VCAA 2017 Exam 1 Data analysis Q5



76% of students answered this type of question correctly.

14. The given bar chart shows the distribution of *wind directions* recorded at a weather station at 9:00 am on each of 214 days in 2011.

According to the bar chart, the percentage of the 214 days on which the wind direction was observed to be east or south-east is closest to

A. 10%

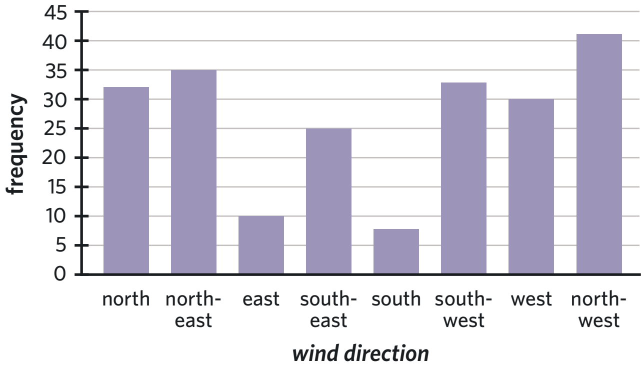
B. 16%

C. 25%

D. 33%

E. 35%

VCAA 2012 Exam 1 Data analysis Q2



68% of students answered this question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

15. The *clothing size* (small, medium, large), and *age* (under 10 years, 10 years or over) of students at a primary school were collected. In this context, the variables *clothing size* and *age* are

A. ordinal and nominal respectively

B. nominal and ordinal respectively

C. ordinal and continuous respectively

D. both ordinal

E. both nominal

Adapted from VCAA 2016 Exam 1 Data analysis Q2

Recursion and financial modelling *Year 11 content*

16. As part of her new year resolutions, Sarah decides to read every month from January to December for one year. Each month she counts the number of pages that she has read.

In January, she reads 12 pages of a book. In February, she reads 18 pages. In March, she reads 24 pages. In April, she reads 30 pages.

The number of pages she reads each month continues to increase according to this pattern.

The number of pages she reads in September is

A. 48

B. 54

C. 60

D. 66

E. 72

Adapted from VCAA 2014 Exam 1 Number patterns Q1

Recursion and financial modelling *Year 11 content*

17. Alex invested $1000 in a savings account, with interest compounding annually.

M_n is the amount of money in the account after n years.

The following calculations show the amount of money in Alex's account initially, and after one and two years.

$M_0 = 1000$

$M_1 = 1.04 \times 1000 = 1040$

$M_2 = 1.04 \times 1040 = 1081.60$

a. Find a recurrence relation in terms of M_0, M_{n+1}, and M_n that models the amount of money in Alex's savings account after n years. (1 MARK)

b. Alex wants to buy a new laptop for $1250. What is the minimum interest rate per annum that would have been required for Alex to afford this laptop after two years? Give your answer correct to two decimal places. (1 MARK)

Adapted from VCAA 2018NH Exam 2 Recursion and financial modelling Q7c,d

1C Displaying numerical data

STUDY DESIGN DOT POINT

- representation, display and description of the distributions of numerical variables: dot plots, stem plots, histograms; the use of a logarithmic (base 10) scale to display data ranging over several orders of magnitude and their interpretation in terms of powers of ten

1A 1B **1C** 1D 1E 1F 1G 1H 1I

KEY SKILLS

During this lesson, you will be:

- displaying data using dot plots
- displaying data using stem plots
- constructing grouped frequency tables
- displaying data using histograms.

KEY TERMS

- Dot plot
- Stem plot
- Grouped frequency table
- Histogram

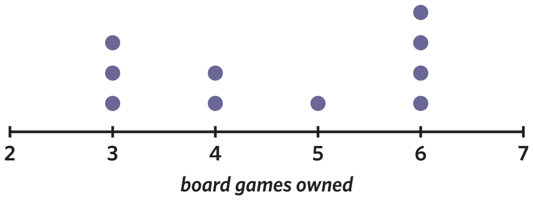
Dot plots, stem plots and histograms are displays that help us visualise the distribution of numerical data. These displays can then be used to identify the number, or percentage, of data within certain ranges of values, as well as the most frequently occurring values.

Displaying data using dot plots

A **dot plot** is a simple way to display discrete numerical data, where each data point is represented by a dot above a single axis.

The number of dots above a value on the axis represents the frequency of the value. The mode of the data set (also known as the modal value) is the value with the most number of dots.

Dot plots are ideal for displaying small/medium-sized data sets with a small range of values.



Worked example 1

Sophie surveyed 12 of the families living on her street.

She asked for the *number of pets* each of them owned and the results were recorded.

3 2 0 1 1 3 0 5 2 1 1 2

a. Construct a dot plot to display this data.

Explanation

Step 1: Rearrange the data set into ascending order and determine the lowest and highest value.

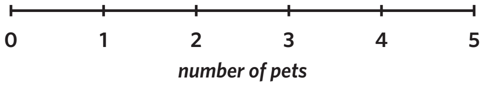
0 0 1 1 1 1 2 2 2 3 3 5

The lowest value is zero.

The highest value is five.

Step 2: Construct a number line with an appropriate scale.

The scale should cover all values between zero and five.



Continues →

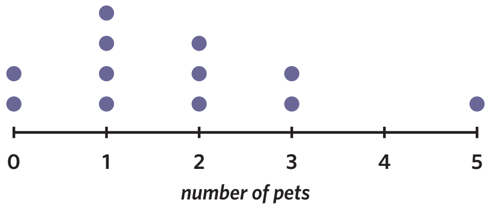
Step 3: Represent each value with a dot.

Mark a dot above the number on the number line each time a value appears in the data set.

If the same data value appears multiple times, illustrate this by placing the corresponding number of dots in a vertical line.

Spacing between each of the vertical dots should be consistent to allow for comparison of frequency across different values.

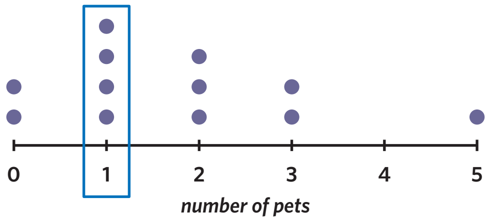
Answer



b. What was the modal *number of pets* owned by families that Sophie surveyed?

Explanation

Find the value with the most dots on the dot plot.



Answer

1 pet

Displaying data using stem plots

See worked example 2

A **stem plot** is a way to display numerical data, where data points are grouped by their leftmost digit(s). Each leaf represents the last digit of an individual data value, and each stem represents the leftmost digit(s) of a group of leaves.

Stems are shown vertically, to the left of a vertical line, ordered from smallest to largest.

leaves are positioned to the right of the vertical line, in line with their corresponding stem. Within each stem, leaves should be ordered from smallest to largest.

Key: 4 | 3 = 43

4	3	5	5	5	7
5	1	3	6	8	
6	7	9			
7	1				
8	0	1	3	9	9
9	1	4			

When constructing a stem plot, always remember to include a key. The key demonstrates the scale of the data. The key allows for data of many forms to be shown in a stem plot. This includes decimals and three (or more) digit numbers.

Decimal:

Key: 1 | 2 = 1.2

1	3 3 4 6 8
2	0 4 9
3	1 1 1 4 5 8
4	2

Three-digit:

Key: 20 | 0 = 200

20	0 0 1 1 2
21	0 1 3 5
22	1 6 9
23	0 4 4 8 8 8
24	5 7

The frequency of a single data value can be found by finding the corresponding stem and counting the number of corresponding leaves within it. The modal value is the value with the most number of identical leaves within a single stem.

See worked example 3

In some cases it can be difficult to see the underlying distribution due to having a lot of data within a small range. This problem is solved by 'splitting' the stems. Usually each stem is split into either two or five stems, depending on how close together the data is.

Key: 1 | 2 = 12

0	1 1 2 3 4
0	5 6 6 8 8 9
1	2 3 3
1	6 7 7 8 9 9

Stem plots are ideal for displaying small/medium-sized data sets with a large range of values.

Worked example 2

Ms Smyth's maths class of 25 students sat their end-of-year exam. Their *results* (%) were recorded.

55	68	76	90	83	89	75	66	59	84	48	62	58
95	80	77	61	92	99	63	84	65	70	81	96	

a. Construct a stem plot to display this data.

Explanation

Step 1: Consider the most appropriate scale.

The data values are two-digit numbers.

The stems will refer to 'tens'.

The leaves will refer to 'ones'.

Step 2: Fill in the appropriate stems.

The data values range from 48 to 99.

All values which fall in the 40s, 50s, 60s, 70s, 80s and 90s need to be covered.

The appropriate stems are 4, 5, 6, 7, 8 and 9.

4	
5	
6	
7	
8	
9	

Note: Each stem within the range of the data needs to be included, even if there are no data values within it.

Step 3: Fill in the leaves for each stem.

Start with the smallest stem and fill the corresponding leaves in ascending order.

Repeat this for each stem.

4	8
5	5 8 9
6	1 2 3 5 6 8
7	0 5 6 7
8	0 1 3 4 4 9
9	0 2 5 6 9

Step 4: Construct a key.

A key shows the scale in which the data is represented.

As decided in step 1, the stems refer to 'tens' and the leaves refer to 'ones'.

Demonstrate this scale with an example.

Continues →

Answer

Key: 4 | 8 = 48%

4	8					
5	5	8	9			
6	1	2	3	5	6	8
7	0	5	6	7		
8	0	1	3	4	4	9
9	0	2	5	6	9	

b. How many students scored above 70% on the exam?

Explanation

Count the number of leaves that represent a value greater than 70.

This will include any leaves on the '7' stem that are greater than 0 and all leaves on stems greater than 7.

Key: 4 | 8 = 48%

4	8					
5	5	8	9			
6	1	2	3	5	6	8
7	0	5	6	7		
8	0	1	3	4	4	9
9	0	2	5	6	9	

Answer

14 students

Worked example 3

Ms Goyle's maths class of 25 students sat their end-of-year exam. Their *results* (%) were recorded.

75	68	76	80	83	69	65	66	79	84	78	62	88
75	80	77	61	62	69	73	84	75	60	81	66	

Construct a split stem plot to display this data, with stem intervals of 5%.

Explanation

Step 1: Consider the most appropriate scale.

The data values are two-digit numbers.

The stems will refer to 'tens'.

The leaves will refer to 'ones'.

Step 2: Fill in the appropriate stems.

The data values range from 60 to 88.

All values which fall in the 60s, 70s, and 80s need to be covered.

The question specifies stem intervals of 5%.

The appropriate stems are 6, 6, 7, 7, 8 and 8.

6	
6	
7	
7	
8	
8	

Continues →

Step 3: Fill in the leaves for each stem.

Start with the smallest stem and fill the corresponding leaves in ascending order.

The top stem for each stem value will include leaves from 0–4 and the bottom stem for each value will include leaves from 5–9.

Repeat this for each stem.

6	0	1	2	2			
6	5	6	6	8	9	9	
7	3						
7	5	5	5	6	7	8	9
8	0	0	1	3	4	4	
8	8						

Step 4: Construct a key.

A key shows the scale in which the data is represented.

As decided in step 1, the stems refer to 'tens' and the leaves refer to 'ones'.

Demonstrate this scale with an example.

Answer

Key: 6 | 0 = 60%

6	0	1	2	2			
6	5	6	6	8	9	9	
7	3						
7	5	5	5	6	7	8	9
8	0	0	1	3	4	4	
8	8						

Constructing grouped frequency tables

A **grouped frequency table** groups data in regular intervals, and displays the frequency and percentage frequency of each interval. This allows the distribution of the data to be more clearly observed.

The lower bound of each interval is inclusive, whilst the upper bound is not.

For example, this grouped frequency table demonstrates that 46 students scored at least 70% but less than 80% on the test.

test mark	**frequency**	
	number	**%**
50–<60%	12	10.0
60–<70%	35	29.2
70–<80%	46	38.3
80–<90%	19	15.8
90–<100%	8	6.7
total	120	100.0

Worked example 4

The *height* (cm) of 20 plants in a Year 6 science experiment were recorded.

32.0	40.2	40.5	45.1	47.0	49.1	50.1	53.7	54.2	55.3
56.9	57.2	58.2	67.2	68.9	69.0	72.3	77.6	82.1	88.5

Construct a grouped frequency table with intervals of 10.

Explanation

Step 1: Determine the number of intervals required.

The question specifies intervals of 10.

The data has a minimum of 32.0, so the intervals should start at 30.

The data has a maximum of 88.5, so the intervals should end at 90.

Therefore, six intervals will be required.

Continues →

Step 2: Set up a table.

The table should have 3 columns, for the variable collected and for the frequency (as a number and a percentage). There should be enough rows to include all the intervals, the header and the total.

Each interval includes the lower bound and does not include the upper bound. For example, the first interval is from 30 to less than 40, and is written as '30–<40'.

height **(cm)**	**frequency**	
	number	**%**
30–<40		
40–<50		
50–<60		
60–<70		
70–<80		
80–<90		
total		

Step 3: Fill in the frequency columns.

Count the number of data values within each interval in the data set. This is displayed in the number column. The total of the number column will be the sum of the frequencies of the intervals.

Then find the percentage frequency for each interval. The total percentage frequency will always be 100.

$$\textit{percentage frequency} = \frac{\textit{frequency}}{\textit{total frequency}} \times 100$$

Answer

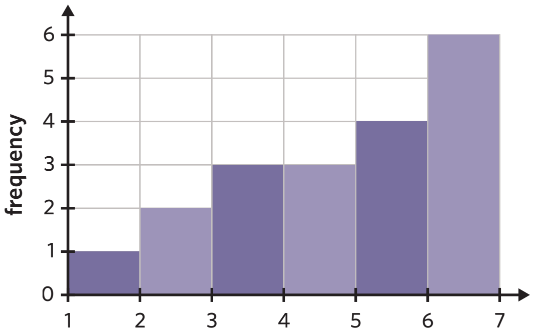
height **(cm)**	**frequency**	
	number	**%**
30–<40	1	5
40–<50	5	25
50–<60	7	35
60–<70	3	15
70–<80	2	10
80–<90	2	10
total	20	100

Displaying data using histograms

A **histogram** is a visual representation of a grouped frequency distribution table. It can display either frequency or percentage frequency on its vertical axis.

See worked example 5

As a histogram depicts intervals of values, there are no spaces between columns.



The height of a column represents the frequency, or percentage frequency, of the interval of values. The modal interval is the interval with the tallest column.

Histograms are best at displaying data sets with a large number of numerical values. A calculator can be helpful in creating a histogram directly from the data.

See worked example 6

Worked example 5

The *height* (cm) of 20 plants in a Year 6 science experiment was recorded in the following grouped frequency table.

height (cm)	frequency	
	number	**%**
30–<40	1	5
40–<50	5	25
50–<60	7	35
60–<70	3	15
70–<80	2	10
80–<90	2	10
total	20	100

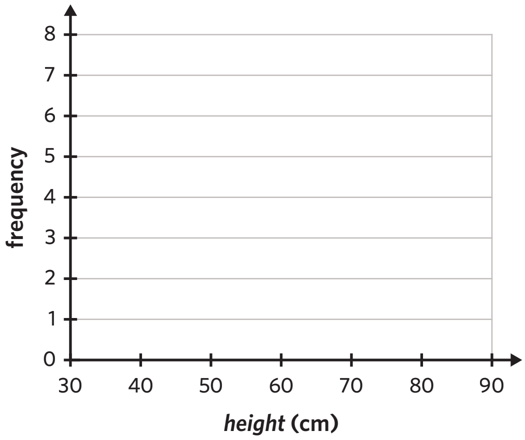
a. Construct a histogram to display this data.

Explanation

Step 1: Construct a set of axes.

Label the horizontal axis '*height* (cm)'. The horizontal axis needs to range from at least 30 to 90, with the boundaries of the intervals labelled.

Label the vertical axis 'frequency'. The vertical axis needs to range from at least 0 to 7.

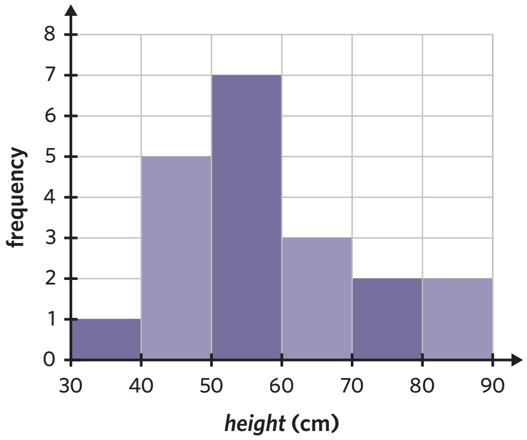


Step 2: Draw a column for each interval.

The height of the column represents the frequency of data values within the corresponding interval.

Histograms do not have spaces between columns.

Answer

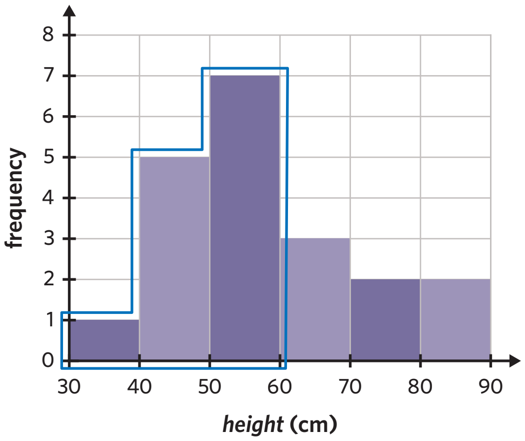


Continues →

b. What percentage of plants were shorter than 60 cm?

Explanation

Step 1: Sum the frequencies of columns with intervals that fall below 60 cm.



1 + 5 + 7 = 13

Step 2: Find this as a percentage of the total number of plants.

There are 20 plants in total.

$$\frac{13}{20} = 0.65 = 65\%$$

Answer

65%

Worked example 6

The *height* (cm) of dogs in a shelter were recorded.

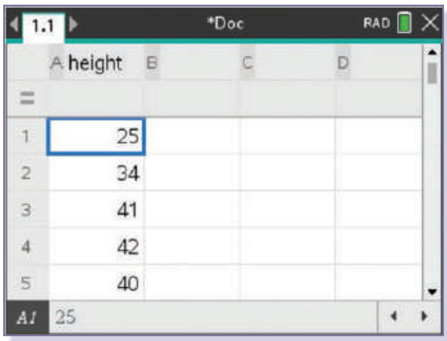
25	34	41	42	40	37	76	47	53	39	72	29	64	61	60
75	39	30	92	45	72	85	50	72	76	42	65	79	64	32

Use the data to construct a histogram.

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Name column A 'height' and enter the data values into column A, starting from row 1.



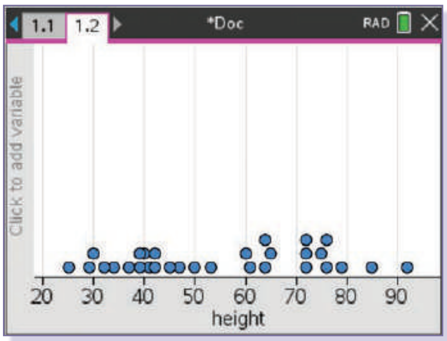
	A height	B	C	D
=				
1	25			
2	34			
3	41			
4	42			
5	40			

A1 25

Step 3: Press ctrl + doc▾ and select '5: Add Data & Statistics'.

Move the cursor to the horizontal axis and select 'Click to add variable'.

Select 'height'.



Step 4: Press menu. Select '1: Plot Type' → '3: Histogram'.

Continues →

Step 5: To adjust the column width and the starting point, press menu and then select ‘2: Plot Properties’ → ‘2: Histogram Properties’ → ‘2: Bin Settings’ → ‘1: Equal Bin Width’.

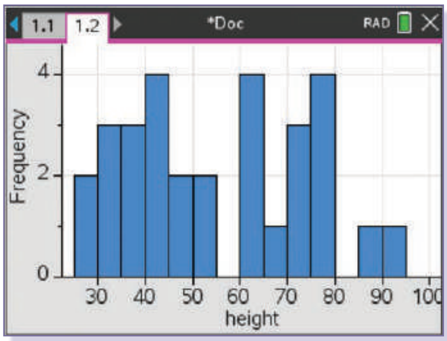
Set the column width to 5 by changing ‘Width’ to ‘5’.

Set the starting point to 25 by changing ‘Alignment’ to ‘25’.

Select ‘OK’.

Note: To change the view of the histogram press menu and use options within ‘5: Window/Zoom’.

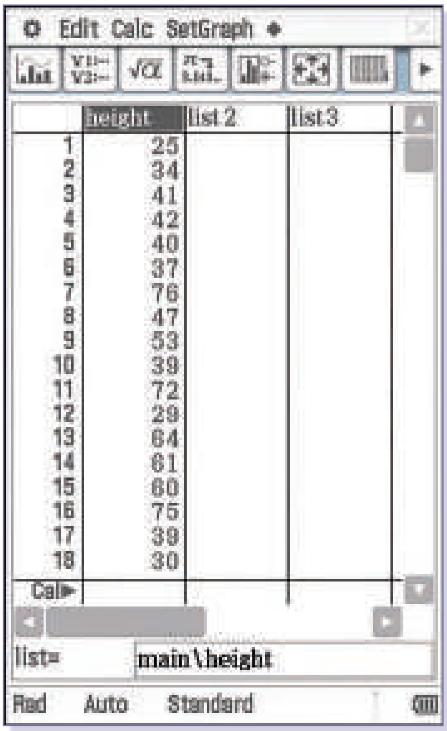
Answer



Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Rename list1 to ‘height’ and enter the data values starting from row 1.



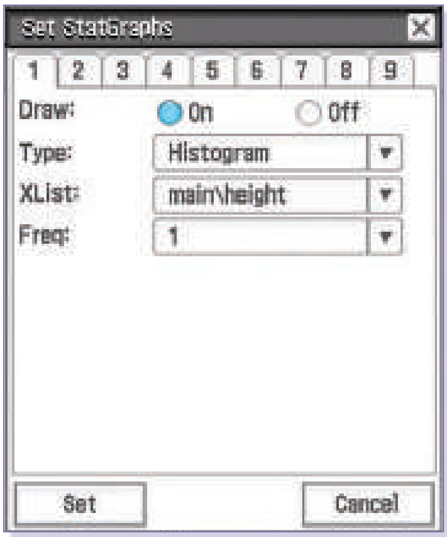
	height
1	25
2	34
3	41
4	42
5	40
6	37
7	76
8	47
9	53
10	39
11	72
12	29
13	64
14	61
15	60
16	75
17	39
18	30

list= main\height

Step 3: Configure the settings of the graph by tapping .

Create a histogram by changing ‘Type:’ to ‘Histogram’.

Specify the data set by changing ‘XList:’ to ‘main\height’.



Draw: ● On ○ Off

Type: Histogram

XList: main\height

Freq: 1

Tap ‘Set’ to confirm.

Step 4: Tap in the icon bar to plot the histogram.

Set the starting point to 25 by changing ‘HStart’ to ‘25’.

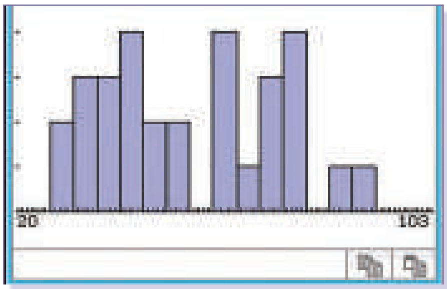
Set the column width to 5 by changing ‘HStep’ to ‘5’.

Tap ‘OK’ to confirm.

Note: To change the view of the histogram press .

Continues →

Answer

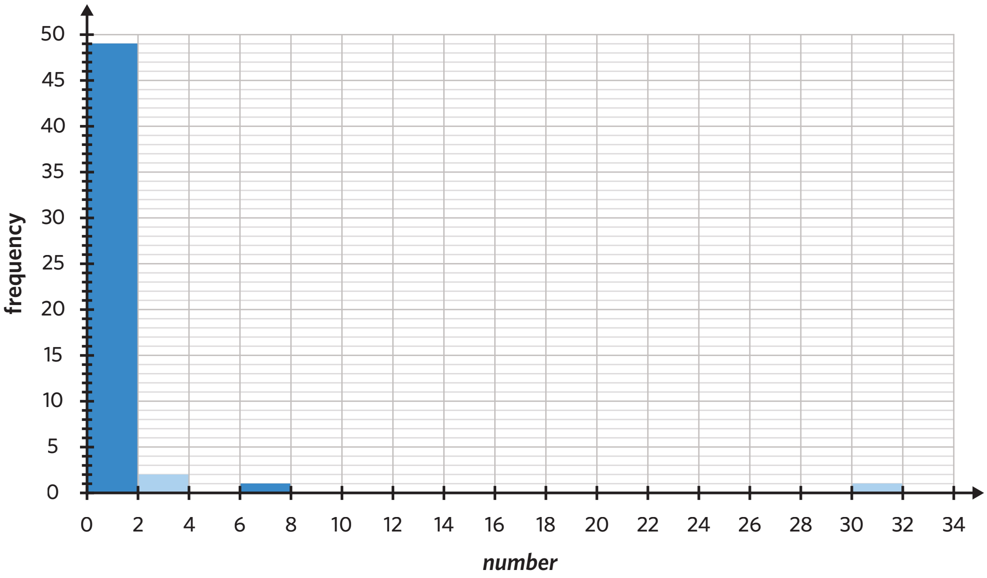


Note: To analyse the histogram, tap in the icon bar to place a (+) marker at the first column. 'xc=25' denotes the starting point of the first column (25) and 'Fc=2' shows the frequency of this interval (2).

Exam question breakdown

VCAA 2016 Exam 1 Data analysis Q6

The following histogram shows the distribution of the *number* of billionaires per million people for 53 countries.



Using this histogram, the percentage of these 53 countries with less than two billionaires per million people is closest to

A. 49% **B.** 53% **C.** 89% **D.** 92% **E.** 98%

Explanation

Step 1: Sum the frequencies of columns with intervals that fall below 2 billionaires per million people.

The only column that falls below 2 billionaires per million people is the leftmost column.

It has a frequency of 49.

Step 2: Find this as a percentage of the total number of countries.

There are 53 countries in total.

$$\frac{49}{53} = 0.9245...$$

$$\approx 92\%$$

71% of students answered this question correctly.

11% of students incorrectly answered A. This was likely because the number of countries with less than 2 billionaires was 49. However, the question asked for this as a percentage of the total number of countries, which was closest to 92%.

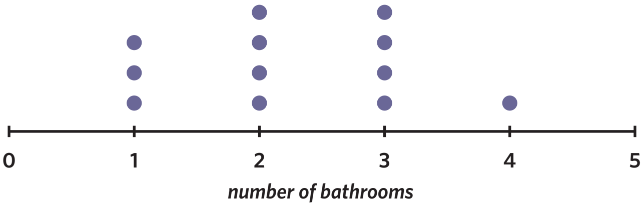
Answer

D

1C Questions

Displaying data using dot plots

1. The *number of bathrooms* in 12 houses is shown in the following dot plot.



The data used to construct this dot plot was

A. 1 2 3 1 2 3 3 2 2 3 1 3

B. 1 3 2 4 2 1 3 3 3 2 1 2

C. 3 3 3 4 4 1 1 3 3 2 2 1

D. 3 1 4 2 2 3 1 1 3 2 2 1

2. Use the following data to construct a dot plot.

21 28 22 23 21 24 27 28 29 24 28 25 25 24

3. Mr Hogan recorded the *amount of time* his students spent over a week studying for history, correct to the nearest hour.

4 3 5 3 2 6 1 2 5 4 3 4 6 4 2 3 2 4 6 3

a. Display Mr Hogan's results in a dot plot.

b. How many students spent two hours studying history over a week?

c. Mr Hogan recommends his students spend 4 hours studying history each week. How many students spent less time than recommended studying history?

d. What percentage of students spent at least five hours studying history over a week?

Displaying data using stem plots

4. The *numbers of fingers* of 11 species of aliens are shown in the following stem plot.

Key: 0 | 5 = 5 fingers

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

The data used to construct this stem plot was

A. 1 23 7 23 5 16 27 8 10 22 31

B. 27 8 10 7 23 5 16 22 31 10 13

C. 23 5 16 22 1 10 13 23 8 13 7

D. 1 13 8 10 22 31 7 23 5 16 27

5. Use the following data to construct a stem plot.

40 89 75 86 89 54 87 82 46 41

74 44 63 71 53 58 91 99 54 47

6. Simon conducted a survey for his statistics class by asking 25 of his classmates how long it took them to get to school on a particular day. He displayed his findings in a stem plot.

Key: 0 | 5 = 5 minutes

0	5 5 6
1	2 3 6 8 9
2	1 4 5 5 6 7 9
3	2 3 4 8 8 8
4	0 1 6
5	2

a. How many students took between 20 and 30 minutes to get to school?

b. How many students took more than 25 minutes to get to school?

c. What is the mode?

7. Mrs Jones' physics class did an experiment investigating the *height* that a ball would bounce after being dropped from a ledge. They conducted 20 trials of the experiment and recorded the *heights* (cm) that the ball reached in each of the trials.

178	184	171	180	183	175	189	174	179	184
187	176	170	188	185	190	195	177	181	182

a. Display this data in a split stem plot, with stem intervals of 5 cm.

b. How many times did the ball bounce between 175 cm and 179 cm (inclusive)?

c. How many times did the ball bounce at least 178 cm?

d. What percentage of balls bounced less than 180 cm?

Constructing grouped frequency tables

8. The *weight* of puppies in a litter are recorded in the following grouped frequency table.

weight **(kg)**	**frequency**
0.8–<0.9	2
0.9–<1.0	4
1.0–<1.1	7
1.1–<1.2	3
total	16

a. How many puppies weigh less than 1 kg?

A. 2

B. 4

C. 6

D. 14

b. What percentage of puppies weigh 0.9 kg or more?

c. What is the modal weight interval?

9. The *maximum flying height* (m) of 16 drones was recorded.

15.3	14.5	15.1	15.2	15.0	16.2	14.8	14.9
15.8	16.2	15.9	16.2	15.3	16.1	15.6	14.9

a. Construct a grouped frequency distribution table with an interval size of 0.5 m to display the data.

b. The company claims that its drones can reach a *maximum flying height* of at least 15 m. What percentage of drones do not meet this criterion?

Displaying data using histograms

10. The number of *games played* by 22 Narrm Demons players is shown in the histogram.

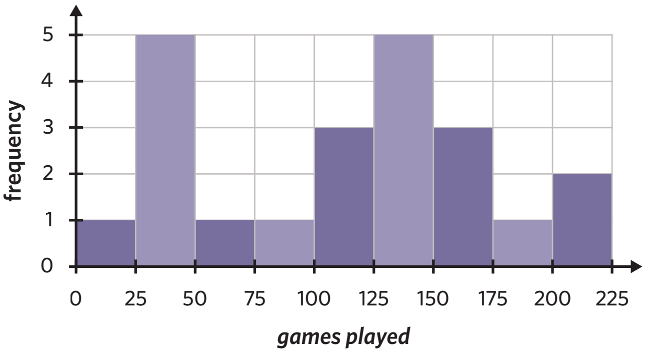
How many players have played less than 100 games?

A. 1

B. 8

C. 11

D. 14



11. A class of 20 students timed themselves to see how long it would take each of them to run around the school oval. Their results are displayed in a histogram.

a. How many students took between

i. 40 and 45 seconds?

ii. 55 and 60 seconds?

iii. 40 and 60 seconds?

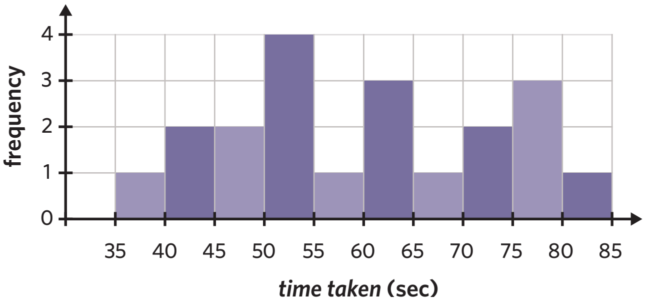
b. How many students ran around the oval in

i. less than 40 seconds?

ii. less than 60 seconds?

iii. more than 70 seconds?

c. What is the modal interval?



12. The *selling prices* of NFTs are shown in the following grouped frequency table.

a. Use the grouped frequency table to construct a histogram.

b. How many NFTs sold for at least $35 000 but less than $55 000?

c. Use the grouped frequency table to construct a histogram using percentage frequency.

d. What percentage of the NFTs sold for at least $40 000, correct to the nearest percent?

selling price ($000's)	frequency	
	number	%
15–<20	2	6.5
20–<25	1	3.2
25–<30	3	9.7
30–<35	7	22.6
35–<40	3	9.7
40–<45	5	16.1
45–<50	1	3.2
50–<55	4	12.9
55–<60	2	6.5
60–<65	3	9.7
total	31	100.0

13. The *ages* of 20 guests at a wedding are recorded.

19 27 45 25 28 55 22 56 93 12 52 88 39 20 18 43 82 45 76 58

a. Use a calculator to construct a histogram. The first column should start at 10 and it should have a column width of 10.

b. Why was 10 chosen as the start of the first column?

c. What is the starting point for the fifth column?

d. What is the modal interval?

Joining it all together

14. The *maximum temperature* (°C) for days within a fortnight in Bairnsdale are listed.

12 15 10 21 35 42 32 28 31 16 18 8 12 10

a. Which visual display is most appropriate for the data?

b. Construct the most appropriate visual display for the data.

c. On how many days was the maximum temperature 10 °C or colder?

d. On what percentage of the days was the maximum temperature greater than 16 °C?

15. The *heights* (m) of the players in a junior level basketball team were recorded.

1.51 1.51 1.63 1.64 1.65 1.67 1.68 1.70 1.75
1.76 1.80 1.85 1.86 1.87 1.89 1.90 1.99

a. Construct a grouped frequency distribution table with an interval size of 0.1. Round percentages to one decimal place.

b. Use the grouped frequency table to construct a histogram.

c. How many members have a height of less than 1.8 m?

d. What is/are the mode(s) of the original data?

e. What is/are the modal interval(s)?

f. Is the mode of the original data a better measure of centre than the modal interval? Explain your answer.

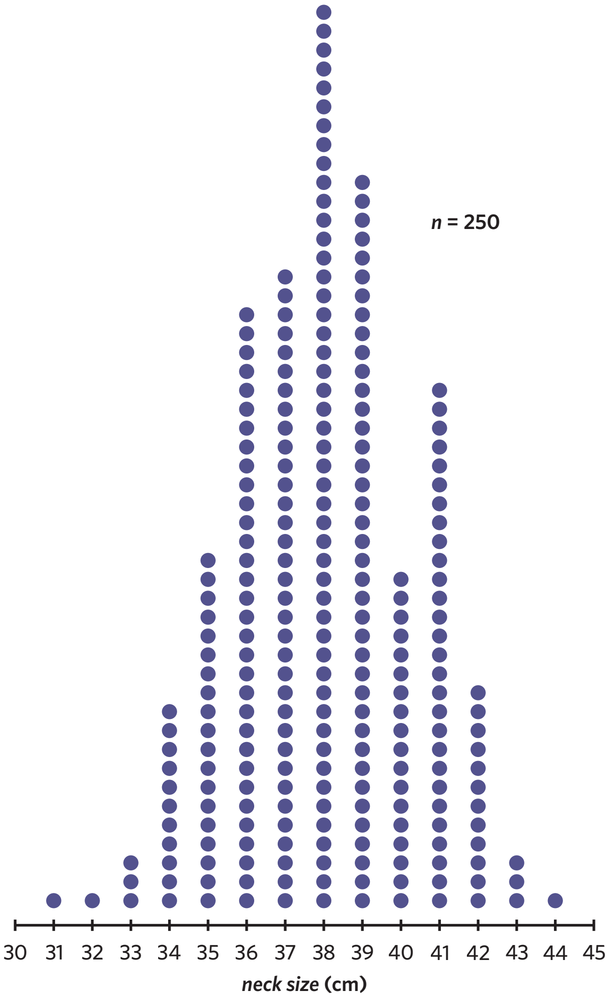
Exam practice

16. The *neck size*, in centimetres, of 250 men was recorded and displayed in the following dot plot.

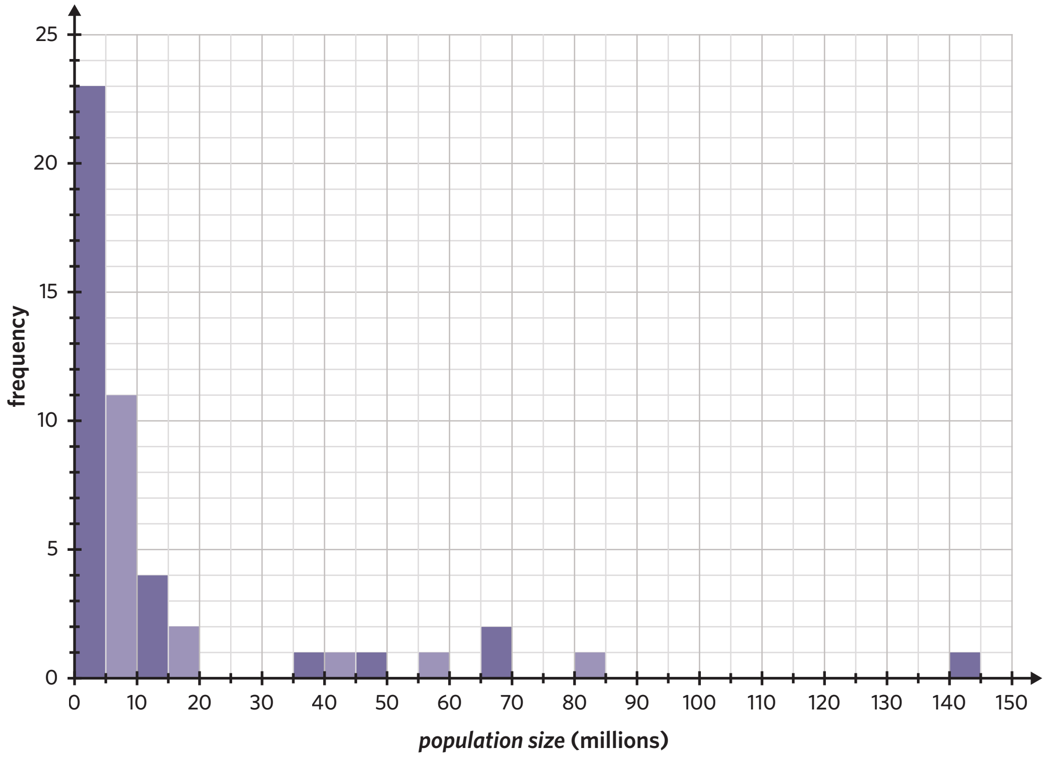
Write down the modal *neck size*, in centimetres, for these 250 men. (1 MARK)

VCAA 2020 Exam 2 Data analysis Q2a

93% of students answered this question correctly.



17. The following histogram shows the distribution of the *population size* of 48 countries in 2018.



The number of these countries with a *population size* between 5 million and 20 million is

A. 11

B. 17

C. 23

D. 34

E. 35

VCAA 2019 Exam 1 Data analysis Q1

92% of students answered this question correctly.

18. The following stem plot displays 30 temperatures recorded at a weather station.

temperature **Key:** 2 | 2 = 2.2 °C

2	2 2 4 4
2	5 7 8 8 8 8 8 8 9 9 9 9
3	1 2 3 3 4 4 4
3	5 6 7 7 7 7
4	1

The modal *temperature* is

A. 2.8 °C

B. 2.9 °C

C. 3.7 °C

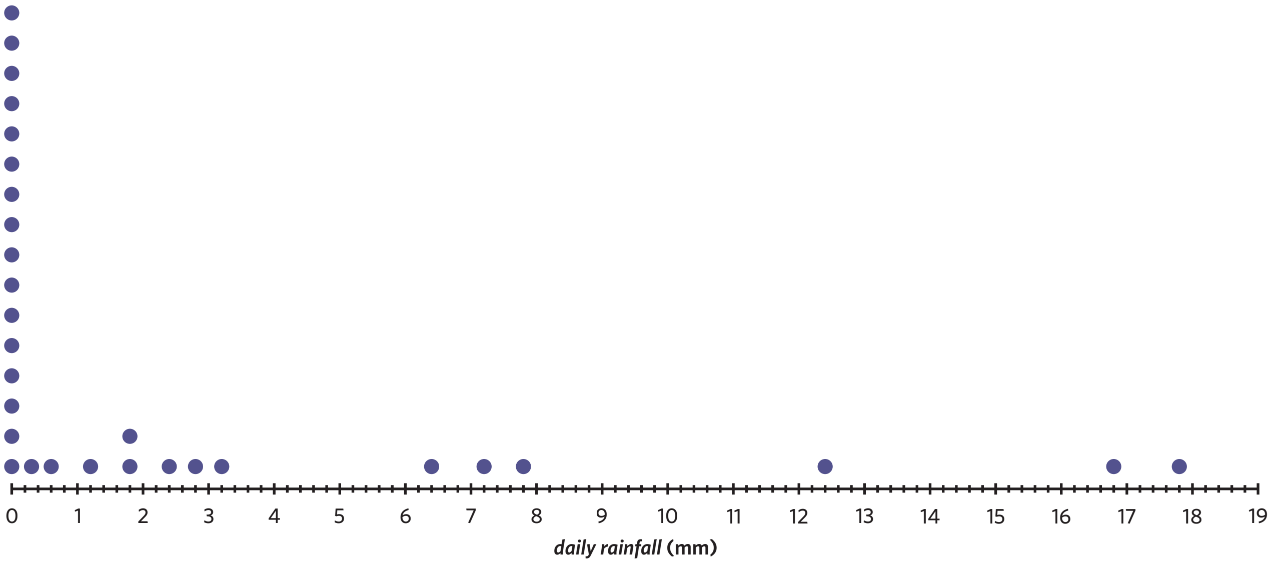
D. 8.0 °C

E. 9.0 °C

VCAA 2016 Exam 1 Data analysis Q3

87% of students answered this question correctly.

19. The following dot plot shows the distribution of *daily rainfall*, in millimetres, at a weather station for 30 days in September.



Construct a histogram that displays the distribution of *daily rainfall* for the month of September. Use interval widths of two with the first interval starting at 0. (2 MARKS)

Adapted from VCAA 2016 Exam 2 Data analysis Q1d

The average mark on this type of question was **1.1**.

Questions from multiple lessons

Data analysis *Year 11 content*

20. The following two-way frequency table displays the *favourite type of pizza* (margherita, capricciosa, hawaiian) and *sex* (male, female) of 140 people.

What percentage of females chose capricciosa as their favourite type of pizza, correct to the nearest percent?

A. 23%

B. 29%

C. 32%

D. 43%

E. 45%

favourite type of pizza	***sex***	
	male	**female**
margherita	12	28
capricciosa	29	32
hawaiian	24	15
total	65	75

Adapted from VCAA 2016 Exam 1 Data analysis Q1

Data analysis *Year 11 content*

21. The heights, in centimetres, of a sample of eight basketball players were recorded and are displayed in the following table.

***height* (cm)**	189	178	190	183	181	194	186	188

The mean, $\bar{x}$, and standard deviation, s_x, of the heights for this sample are closest to

A. $\bar{x} = 186.1 \quad s_x = 4.88$

B. $\bar{x} = 186.1 \quad s_x = 5.00$

C. $\bar{x} = 186.1 \quad s_x = 5.22$

D. $\bar{x} = 4.88 \quad s_x = 186.1$

E. $\bar{x} = 5.22 \quad s_x = 186.1$

Adapted from VCAA 2017 Exam 1 Data analysis Q3

Data analysis *Year 11 content*

22. Information on 26 different animals is shown in a table.

The four variables in this data set are:

- *species* – name of the animal
- *continent of origin* – continent the animal is originally from
- *size* – size of the animal (small, medium, large)
- *average lifespan* – the average lifespan of the animal (years)

species	***continent of origin***	***size***	***average lifespan* (years)**
aardvark	Africa	medium	30
baboon	Africa	large	45
capybara	South America	medium	10
dodo	Africa	medium	20
echidna	Oceania	small	10
ferret	Europe	small	5
goat	Asia	medium	15
hippopotamus	Africa	large	50
impala	Africa	large	12
jaguar	South America	large	15
kangaroo	Oceania	large	20
lemur	Africa	small	16
meerkat	Africa	small	13
nightingale	Europe	small	2
ocelot	South America	medium	10
penguin	Antarctica	medium	20
quokka	Oceania	small	5
red panda	Asia	small	12
scorpion	Africa	small	4
Tasmanian devil	Oceania	medium	5
uakari	South America	medium	15
vulture	North America	medium	20
wallaby	Oceania	medium	9
x-ray tetra	South America	small	4
yak	Asia	large	20
zebra	Africa	large	25

a. How many variables in this data set are categorical? (1 MARK)

b. How many variables in this data set are ordinal? (1 MARK)

c. List the medium-sized animals from Africa. (1 MARK)

Adapted from VCAA 2018 Exam 2 Data analysis Q1

1D Log scales and graphs

STUDY DESIGN DOT POINTS

- representation, display and description of the distributions of numerical variables: dot plots, stem plots, histograms; the use of a logarithmic (base 10) scale to display data ranging over several orders of magnitude and their interpretation in terms of powers of ten
- use of the distribution(s) of one or more categorical or numerical variables to answer statistical questions

1A 1B 1C 1D 1E 1F 1G 1H 1I

KEY SKILLS

During this lesson, you will be:

- calculating logarithmic values
- displaying data using a logarithmic scale
- interpreting data displayed using a logarithmic scale.

KEY TERMS

- Logarithms
- Logarithmic scale

There are times in which the range of data collected is very large and is not feasible to plot on a graph. When data involves values from multiple orders of magnitude (i.e. 1, 10, 100, 1000), it can be difficult to plot them on the same set of axes and still be able to see the distribution clearly. Logarithms can convert these values into smaller numbers. For example, the Richter scale measures the amplitude of earthquakes, which are often very large values. Calculating the log returns a magnitude value, which can then be plotted on a graph.

Calculating logarithmic values

Logarithms, commonly referred to as logs, are a mathematical operation. Each logarithm has a base, an argument, and an exponent.

An expression written in logarithmic form can be written in exponent form and vice versa.

If...

$$\log_{\underset{\textbf{base}}{b}}(\overset{\textbf{argument}}{x}) = y \leftarrow \textbf{exponent}$$

then...

$$b^y = x$$

For example, $\log_{10}(1000) = 3$ since $10^3 = 1000$. In this case, the base is 10, the exponent is 3 and the argument is 1000.

This course only focuses on logarithms with a base of 10. If there is no base specified, then it can be assumed that the base is 10.

If $x > 1$, then $\log(x)$ is positive.

If $x = 1$, then $\log(x)$ is zero.

If $0 < x < 1$, then $\log(x)$ is negative.

If $x \leq 0$, then $\log(x)$ is undefined.

Worked example 1

Use $\log_{10}$ for the following calculations.

a. Calculate the log of 23, correct to two decimal places.

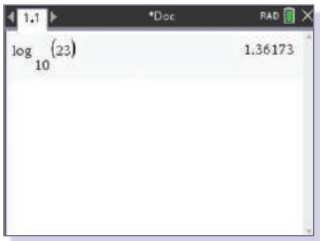
Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select ‘1: New’ → ‘1: Add Calculator’.

Step 2: Press ctrl + 10^x for the logarithmic function.

Enter in the base, ‘10’, and the argument, ‘23’.

Press enter .



$\log_{10}(23)$ 1.36173

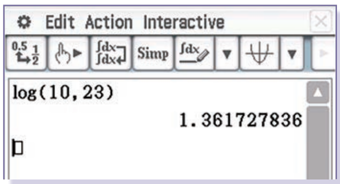
Step 3: Read the value from the screen and round to two decimal places.

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap √α Main .

Step 2: Press keyboard , tap $\log_{\square}(\square)$ and enter ‘23’.

Press EXE.



log(10, 23)

1.361727836

Step 3: Read the value from the screen and round to two decimal places.

Answer – Method 1 and 2

1.36

b. If the log of a number is 5, what is the number?

Explanation

Step 1: Define the unknown number as x.

$\log_{10}(x) = 5$

Step 2: Express this in exponent form and solve for x.

If $\log_{10}(x) = 5$, then $x = 10^5$.

$x = 10^5$

$x = 100\ 000$

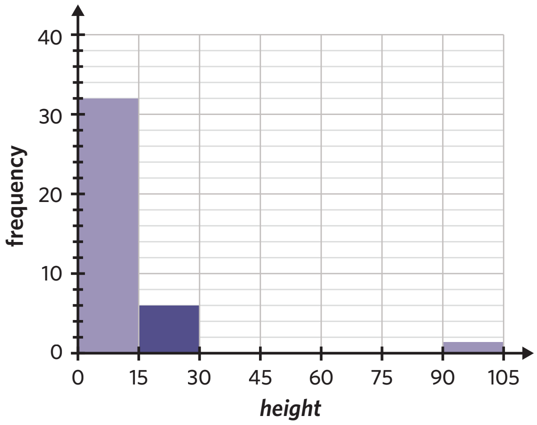
Answer

100 000

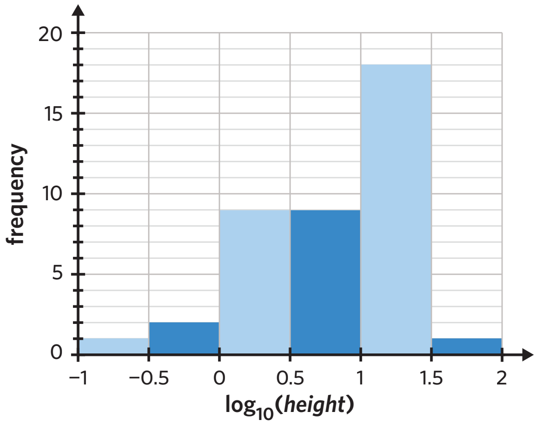
Displaying data using a logarithmic scale

A **logarithmic scale** is a scale with a $\log_{10}$ transformation. It can be used to reveal details about the underlying distribution of data sets with multiple orders of magnitude.

The *height* (cm) of 39 plants are shown on a non-logarithmic scale.



The *height* (cm) of the same plants are shown on a logarithmic scale.



The distribution of values previously hidden within the 0–<15 cm interval is now visible.

Worked example 2

The approximate weights of 20 different vehicles (cars, boats, planes) were recorded in kilograms.

840	910	990	1120	1490	1670	1690	1790	2430	2590
2810	2850	2940	3100	3590	3640	3660	5190	380 000	450 100

Construct a log scale histogram using the data. The histogram should start at 2.9 and have a column width of 0.2.

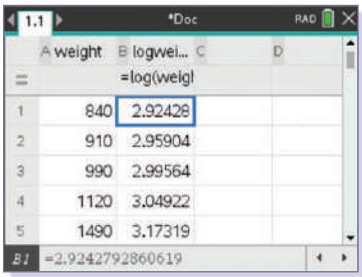
Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select ‘1: New’ → ‘4: Add Lists & Spreadsheet’.

Step 2: Name column A ‘weight’ and enter the data values into column A, starting from row 1.

Step 3: Name column B ‘logweight’.

Enter ‘=log(weight)’ into the cell below the ‘logweight’ heading.



	A weight	B logwei...	C	D
=		=log(weig|		
1	840	2.92428		
2	910	2.95904		
3	990	2.99564		
4	1120	3.04922		
5	1490	3.17319		

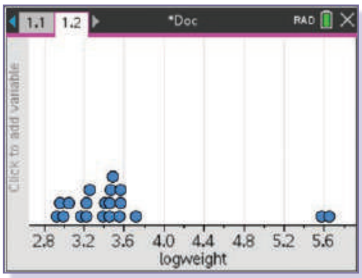
B1 =2.9242792860619

Continues →

Step 4: Press ctrl + doc ▾, and select '5: Add Data & Statistics'.

Move the cursor to the horizontal axis and select 'Click to add variable'.

Select 'logweight'.



Step 5: Press menu. Select '1: Plot Type' → '3: Histogram'.

Step 6: To adjust the column width and the starting point, press menu and then select '2: Plot Properties' → '2: Histogram Properties' → '2: Bin Settings' → '1: Equal Bin Width'.

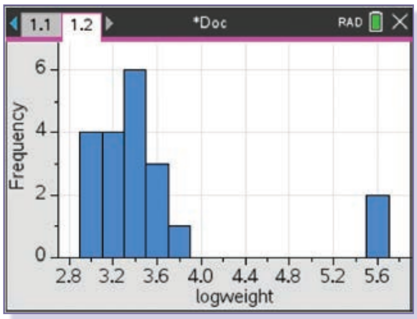
Set the column width to 0.2 by changing 'Width' to '0.2'.

Set the starting point to 2.9 by changing 'Alignment' to '2.9'.

Select 'OK'.

Note: To change the view of the histogram, press menu and use options within '5: Window/Zoom'.

Answer



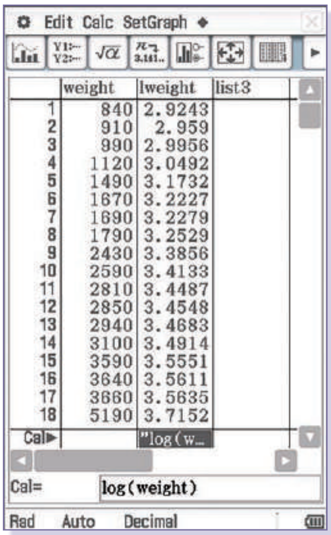
Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Rename list1 to 'weight' and enter the data values starting from row 1.

Step 3: Rename list2 'lweight'.

Go down to the calculation cell Cal▶ and enter 'log(weight)'.



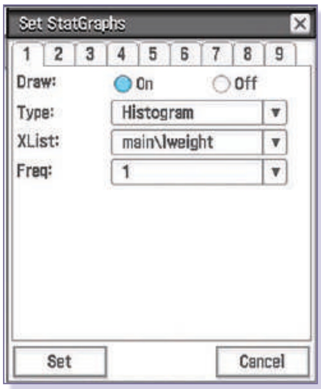
	weight	lweight	list3
1	840	2.9243	
2	910	2.959	
3	990	2.9956	
4	1120	3.0492	
5	1490	3.1732	
6	1670	3.2227	
7	1690	3.2279	
8	1790	3.2529	
9	2430	3.3856	
10	2590	3.4133	
11	2810	3.4487	
12	2850	3.4548	
13	2940	3.4683	
14	3100	3.4914	
15	3590	3.5551	
16	3640	3.5611	
17	3660	3.5635	
18	5190	3.7152	
Cal▶		"log(w…	

Cal= log(weight)

Step 4: Configure the settings of the graph by tapping .

Create a histogram by changing 'Type:' to 'Histogram'.

Specify the data set by changing 'XList:' to 'main\lweight'.



Draw: ● On ○ Off

Type: Histogram

XList: main\lweight

Freq: 1

Tap 'Set' to confirm.

Step 5: Tap in the icon bar to plot the histogram.

Set the starting point to 2.9 by changing 'HStart' to '2.9'.

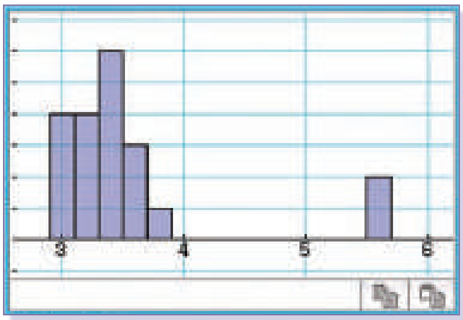
Set the column width to 0.2 by changing 'HStep' to '0.2'.

Tap 'OK' to confirm.

Note: To change the view of the histogram, press .

Continues →

Answer



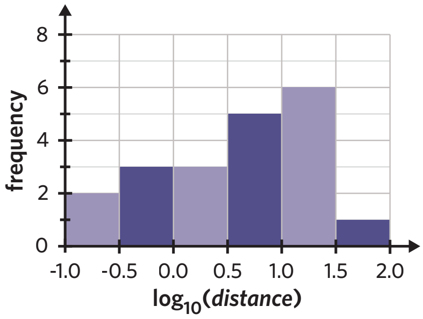
Note: To analyse the histogram, tap in the icon bar to place a (+) marker at the first column. 'xc=2.9' denotes the starting point of the first column (2.9) and 'Fc=4' shows the frequency of this interval (4).

Interpreting data displayed using a logarithmic scale

The height of a column in a histogram displayed using a logarithmic scale continues to represent the frequency of the corresponding interval. However, to interpret the frequency of an interval, it is important to consider the log transformation that has been applied to the horizontal axis.

Worked example 3

A class of twenty students were asked about the *distance*, in kilometres, from their home to the school. The results were summarised in a histogram using a $\log_{10}$ scale.



a. Jane was away the day of the survey. It is known that she lives 9.5 km away from school. Which interval on the histogram reflects this?

Explanation

Step 1: Calculate the appropriate log value.

$\log_{10}(9.5) = 0.9777...$

Step 2: Identify the appropriate interval on the graph.

This would be plotted between 0.5 and 1.0.

Answer

0.5–<1.0

Continues →

b. How many students live more than 10 km away from school?

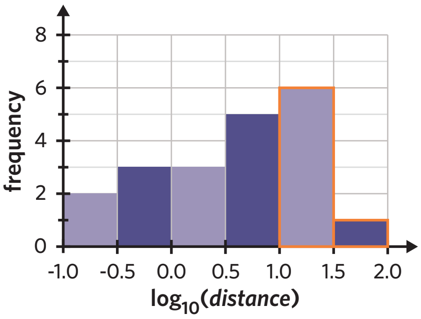
Explanation

Step 1: Calculate the appropriate log value.

$\log_{10}(10) = 1$

This means that on a log scale, 10 is plotted as 1.

Step 2: Sum the columns greater than 1.



$6 + 1 = 7$

Answer

7 students

Exam question breakdown

VCAA 2016 Exam 1 Data analysis Q7

The following histogram shows the distribution of the *number* of billionaires per million people plotted on a $\log_{10}$ scale.

Based on this histogram, the number of countries with one or more billionaires per million people is

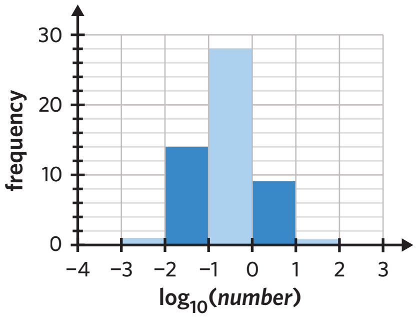
A. 1

B. 3

C. 8

D. 9

E. 10



Data: Gapminder

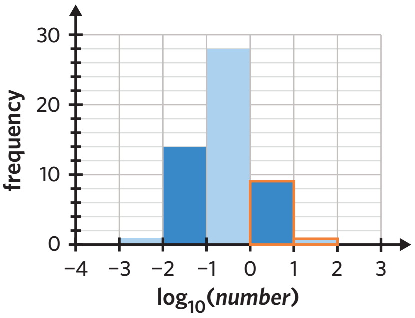
Explanation

Step 1: Calculate the appropriate log value.

$\log_{10}(1) = 0$

This means that on a log scale, 1 is plotted as 0.

Step 2: Sum the columns higher than 0.



$9 + 1 = 10$

Answer

E

45% of students answered this question correctly.

29% of students incorrectly chose option A. These students likely ignored the log scale and counted the column greater than 1, rather than the columns greater than $\log_{10}(1)$.

1D Questions

Calculating logarithmic values

1. What is the value of $\log_{10}(10)$?

A. −1 **B.** 0 **C.** 1 **D.** 10

2. Calculate the log values of the following, correct to two decimal places.

a. 0.05 **b.** 0.5 **c.** 5 **d.** 50

e. 3800 **f.** 380 **g.** 38 **h.** 3.8

3. Calculate the value of x, correct to two decimal places.

a. $\log_{10}(x) = -0.5$ **b.** $\log_{10}(x) = 0.5$ **c.** $\log_{10}(x) = 1.5$

Displaying data using a logarithmic scale

4. Consider the following histogram with a $\log_{10}$ scale.

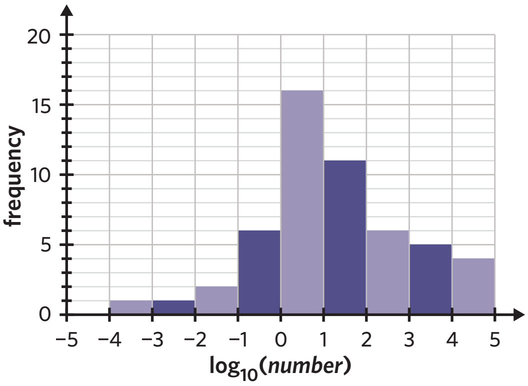
If a student wishes to add the *number* 10 358 to the distribution, which interval would need to be adjusted?

A. 1–<2

B. 2–<3

C. 3–<4

D. 4–<5



5. The following table shows the revenues of six businesses.

***revenue* ($)**	100 000	300 000	135 000	120 000	450 000	175 000
$\log_{10}$(*revenue*)	5.00	5.48	5.13	5.08	5.65	5.24

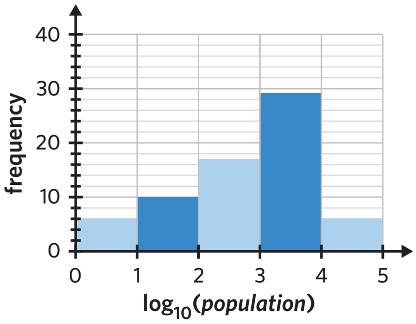
Construct a histogram representing *revenue* with a $\log_{10}$ scale. The histogram should start at 5 and have a column width of 0.1.

6. The *weight* (kg) of animals at the jungle party were recorded. Use a calculator to construct a $\log_{10}$ scale histogram using the data, with an appropriate starting point and a column width of 0.2.

50 56 84 95 87 64 512 204 983 10 200 124 94 43 302 9521

Interpreting data displayed using a logarithmic scale

7. Statisticians have collected data on the *population* of 68 of rural communities. Their data has been displayed as a histogram on a $\log_{10}$ scale.



a. How many communities have a *population* of at least 10 000 people?

A. 0

B. 3

C. 6

D. 62

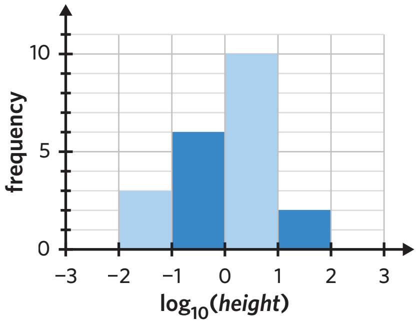
b. The percentage of communities with a *population* of less than 10 people is closest to

A. 6% **B.** 8% **C.** 9% **D.** 11%

8. The *height* (m) of 21 oak trees have been measured by an environmentalist. The data has been displayed as a histogram with a $\log_{10}$ scale.

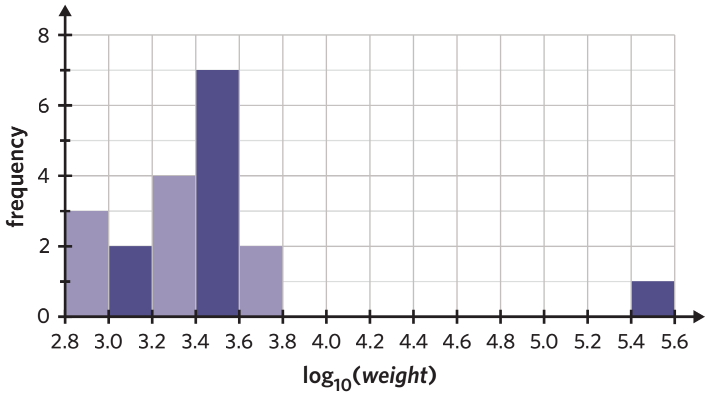
a. How many trees are at least 1 metre tall?

b. What percentage of trees are shorter than 1 metre? Round to two decimal places.



Joining it all together

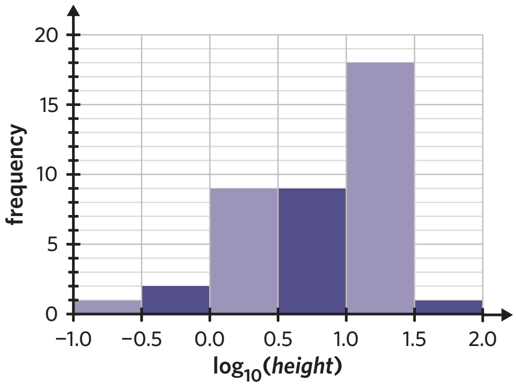
9. The approximate *weight* of different vehicles (cars, boats, planes) were recorded in kilograms. The data was displayed by a histogram using the $\log_{10}$ scale.



a. What *weight* is represented by 3.6 on the log scale, correct to the nearest kg?

b. A truck is known to have a *weight* of 3750 kg. In which interval will this value be plotted?

10. It is the year 3000. Many different alien species have been found across the universe. The $\log_{10}$(*height*), with *height* measured in metres, of a single alien from each species was recorded and displayed in the following histogram.



a. An alien from planet Zigzagzoon is 1.5 m tall. What is its log value, correct to two decimal places?

b. How many aliens are at least 10 m tall?

c. An alien from planet Bazinga is found to have a $\log_{10}$(*height*) value of -0.7, whereas an alien from planet Plazong has a $\log_{10}$(*height*) of 1.6. How much taller (in metres) is the alien from planet Plazong than the alien from planet Bazinga, correct to two decimal places?

11. The weights of 10 vehicles were recorded in kilograms.

a. Using a $\log_{10}$ scale, display the data in the frequency table as a histogram with a column width of 1. Label axes as appropriate.

b. Interpret the value of the modal interval.

vehicle	***weight* (kg)**
bicycle	15
motorcycle	230
truck	6500
van	1500
plane	55 000
hatchback car	1200
sedan car	1300
limousine	2800
scooter	4
tram	32 000

Exam practice

12. The following histogram shows the distribution of *weight*, in grams, for a sample of 20 animal species. The histogram has been plotted on a $\log_{10}$ scale.

The percentage of these small animal species with a *weight* of less than 10 000 g is

A. 17%

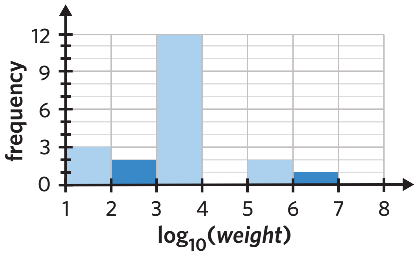
B. 70%

C. 75%

D. 80%

E. 85%

VCAA 2020 Exam 1 Data analysis Q5



78% of students answered this question correctly.

13. The following histogram shows the *population size* for 48 countries plotted on a $\log_{10}$ scale.

Based on this histogram, the number of countries with a *population size* that is less than 100 000 people is

A. 1

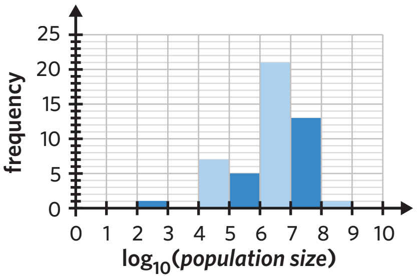
B. 5

C. 7

D. 8

E. 48

VCAA 2019 Exam 1 Data analysis Q3



Data: Worldometers, <www.worldometers.info/>

71% of students answered this question correctly.

14. The following histogram shows the distribution of the $\log_{10}(\textit{area})$ with *area* in square kilometres, of 17 islands.

The modal area of these islands, in square kilometres, is between

A. 2 and 3

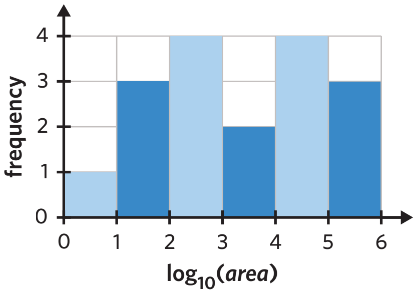
B. 2 and 3, as well as 4 and 5

C. 2 and 5

D. 10 000 and 100 000

E. 100 and 1000, as well as 10 000 and 100 000

Adapted from VCAA 2017 Exam 1 Data analysis Q4

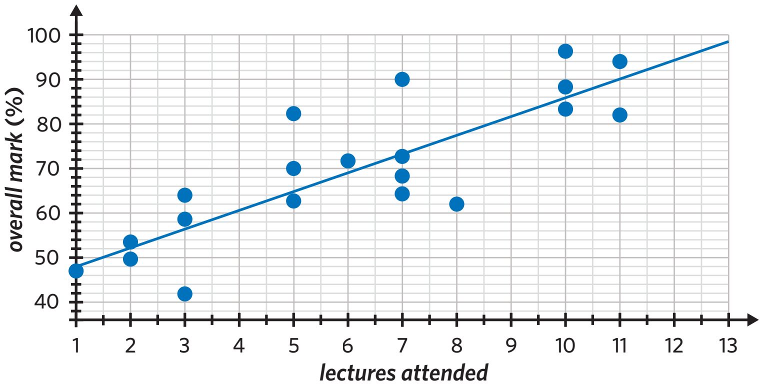


62% of students answered this type of question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

15. The following scatterplot displays the relationship between *overall mark*, as a percentage, and the number of *lectures attended* by 20 students in a particular unit at university. A line of good fit has been drawn.



The equation of the line of good fit is closest to

A. *overall mark* = 43.49 − 4.23 × *lectures attended*

B. *overall mark* = 48.24 + 4.23 × *lectures attended*

C. *overall mark* = 48.24 − 3.98 × *lectures attended*

D. *overall mark* = 43.49 + 3.98 × *lectures attended*

E. *overall mark* = 43.49 + 4.23 × *lectures attended*

Adapted from VCAA 2018 Exam 1 Data analysis Q8

Recursion and financial modelling *Year 11 content*

16. Delores inherits a large sum of money and decides to deposit it in a new savings account which earns interest every year. The following geometric sequence models the value of Delores' investment, in dollars, in each successive year.

850 500, 867 510, 884 860.2, ...

The balance of Delores' savings account after 15 years is closest to

A. $1 105 650

B. $1 108 202

C. $1 110 753

D. $1 122 217

E. $1 144 661

Adapted from VCAA 2013 Exam 1 Number patterns Q4

Recursion and financial modelling *Year 11 content*

17. Denton wants to save up some money to buy a new house. He opens a savings account and deposits some money which will earn compound interest every year.

The balance of Denton's account, in dollars, after *n* years, V_n, can be modelled by the recurrence relation

$V_0 = 87\,500, \quad V_{n+1} = 1.034\,V_n$

a. How many dollars did Denton initially invest? (1 MARK)

b. What is the balance of Denton's account after one year? (1 MARK)

c. How many years will it be until the balance of Denton's account exceeds $100 000? (1 MARK)

Adapted from VCAA 2018 Exam 2 Recursion and financial modelling Q4

1E The five-number summary and boxplots

STUDY DESIGN DOT POINT

- summary of the distributions of numerical variables; the five-number summary and boxplots (including the use of the lower fence ($Q_1 - 1.5 \times IQR$) and upper fence ($Q_3 + 1.5 \times IQR$) to identify and display possible outliers); the sample mean and standard deviation and their use in comparing data distributions in terms of centre and spread

1A 1B 1C 1D **1E** 1F 1G 1H 1I

KEY SKILLS

During this lesson, you will be:

- calculating the five-number summary
- calculating the range and interquartile range
- identifying outliers
- constructing and interpreting boxplots.

KEY TERMS

- Five-number summary
- Minimum
- Maximum
- Median
- Quartiles
- Spread
- Range
- Interquartile range (IQR)
- Outliers
- Fence
- Boxplot

Boxplots show the distribution of a data set based on the five-number summary, instead of displaying the frequency of data values or intervals of data values. The data set is split into quartiles to help visualise the centre and spread of a data set, as well as the existence of any outliers.

Calculating the five-number summary

The **five-number summary** provides key information about a set of data and its distribution including spread and centre. The summary is as follows:

Minimum, Q_1, Median, Q_3, Maximum

The **minimum** is the smallest value in the data set. The **maximum** is the largest value in the data set. It can be helpful to order the data when finding the maximum and minimum values.

The **median** is the middle value in an ordered set of data.

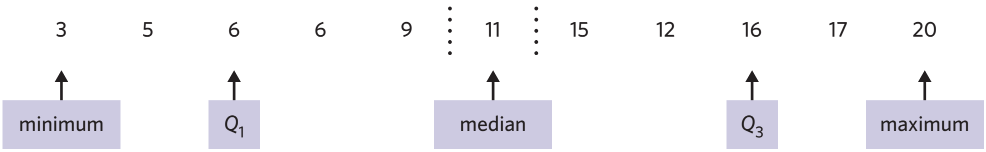
If there are n data values, the median is located at the $\left(\frac{n+1}{2}\right)^{\text{th}}$ position.

When there is an even number of data values, the median will be the average of the two middle data values.

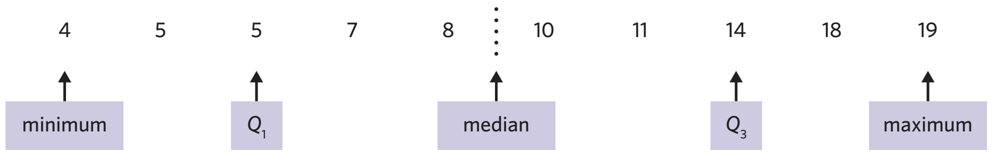
While the median divides a distribution in half, **quartiles** divide a distribution in quarters. The symbols used to refer to the quartiles are Q_1, Q_2 and Q_3.

- Q_1 is the median of the lower half of the data set. The median, Q_2, is excluded if there is an odd number of values.
- Q_2 is the median of the entire data set.
- Q_3 is the median of the upper half of the data set. The median, Q_2, is excluded if there is an odd number of values.

Odd number of values:



Even number of values:



Worked example 1

Construct a five-number summary for the following data.

3 5 1 10 8 9 6 3 8 6

Explanation – Method 1: By hand

Step 1: Arrange the data in ascending order.

1 3 3 5 6 6 8 8 9 10

Step 2: Identify the minimum and maximum values.

$minimum = 1$

$maximum = 10$

Step 3: Determine the median.

Count the number of values in the data set.

$n = 10$

Position of median: $\frac{10 + 1}{2} = 5.5$

The median is the average of the 5th and 6th values.

$median = \frac{6 + 6}{2} = 6$

Step 4: Determine the value of Q_1 and Q_3.

Q_1 is the median of the lower half of the data set.

Lower half: 1 3 3 5 6

$Q_1 = 3$

Q_3 is the median of the upper half of the data set.

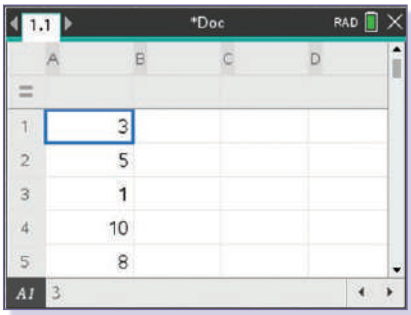
Upper half: 6 8 8 9 10

$Q_3 = 8$

Explanation – Method 2: TI-Nspire

Step 1: From the home screen, select ‘1: New’ → ‘4: Add Lists & Spreadsheet’.

Enter the data values into column A, starting from row 1.



	A	B	C	D
=				
1	3			
2	5			
3	1			
4	10			
5	8			

A1 3

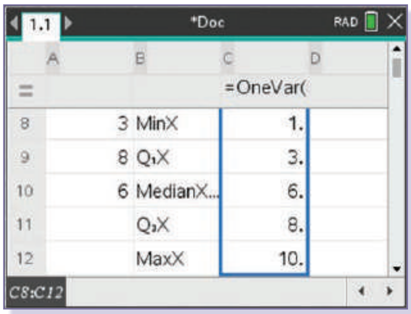
Step 2: Press menu. Select ‘4: Statistics’ → ‘1: Stat Calculations’ → ‘1: One-Variable Statistics’.

Select ‘OK’ to confirm one-variable statistics for one data set only.

Step 3: Specify the data set by entering 'a[]' in ‘X1 List:’.

Select ‘OK’ to exit this window and generate the statistics.

Scroll down to find the five-number summary statistics.



	A	B	C	D
=			=OneVar(	
8	3	MinX	1.	
9	8	Q_1X	3.	
10	6	MedianX...	6.	
11		Q_3X	8.	
12		MaxX	10.	

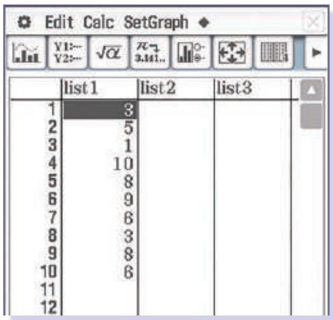
C8:C12

Continues →

Explanation - Method 3: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Enter the data values into list1, starting from row 1.



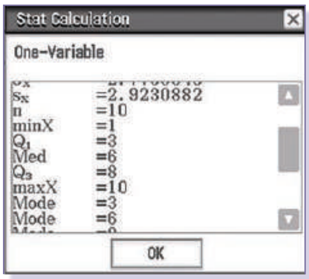
	list1	list2	list3
1	3		
2	5		
3	1		
4	10		
5	8		
6	9		
7	6		
8	3		
9	8		
10	6		
11			
12			

Step 2: Tap ‘Calc’ → ‘One-Variable’.

Specify the data set by keeping ‘XList:’ as 'list1'.

Tap ‘OK’ to confirm.

Step 3: Scroll down to find the five-number summary statistics.



```
One-Variable
Sx      =2.9230882
n       =10
minX    =1
Q1      =3
Med     =6
Q3      =8
maxX    =10
Mode    =3
Mode    =6
```

Answer - Method 1, 2 and 3

1, 3, 6, 8, 10

Calculating the range and interquartile range

The **spread** of a data set refers to how variable or similar the values are.

The **range** is a measure of spread of an entire data set. It is the difference between the maximum and minimum values, even if they are outliers.

$range = maximum - minimum$

The **interquartile range (IQR)** is a measure of the spread of the middle 50% of a data set. It is sometimes more accurate as a measure of spread than range because it does not include the upper 25% and lower 25% of values. This means that it is rarely affected by outliers.

$IQR = Q_3 - Q_1$

Worked example 2

The following stem plot displays the number of *goals scored* by 16 netball teams over a season.

Key: 2 | 2 = 22 goals

2	2 3 5 7 9
3	0 1 3 5 7 7 8
4	0 1 2 2

a. Calculate the range of *goals scored*.

Explanation

Step 1: Identify the minimum and maximum values.

Key: 2 | 2 = 22 goals

2	2 3 5 7 9
3	0 1 3 5 7 7 8
4	0 1 2 2

$minimum = 22$

$maximum = 42$

Step 2: Calculate the range.

$$\begin{aligned} range &= maximum - minimum \\ &= 42 - 22 \\ &= 20 \end{aligned}$$

Answer

20 goals

Continues →

b. Calculate the IQR of *goals scored*.

Explanation

Step 1: Determine the lower half and upper half of the data set.

$n = 16$

Position of median: $\frac{16+1}{2} = 8.5$

Key: 2 | 2 = 22 goals

2	2	3	5	7	9		
3	0	1	3	5	7	7	8
4	0	1	2	2			

Lower half: 22 23 25 27 29 30 31 33

Upper half: 35 37 37 38 40 41 42 42

Step 2: Determine the value of Q_1 and Q_3.

Q_1 is the median of the lower half of the data set.

$Q_1 = \frac{27+29}{2} = 28$

Q_3 is the median of the upper half of the data set.

$Q_3 = \frac{38+40}{2} = 39$

Step 3: Calculate the IQR.

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 39 - 28 \\ &= 11 \end{aligned}$$

Answer

11 goals

Identifying outliers

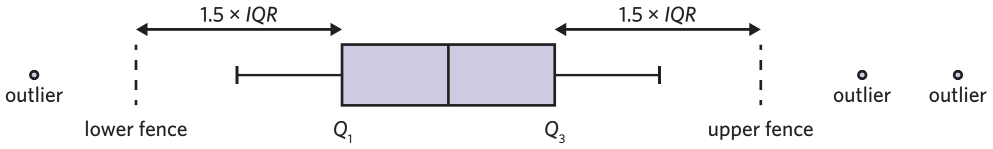
Outliers are values which fall outside of what is normal or reasonable. Outliers can be identified by a series of tests.

A **fence** defines the boundary of what is an outlier, and what is a regular data value.

$\textit{lower fence} = Q_1 - (1.5 \times IQR)$

$\textit{upper fence} = Q_3 + (1.5 \times IQR)$

If a value is less than the lower fence or greater than the upper fence, it is considered to be an outlier. Outliers can still be reported as the maximum or minimum value of a data set.



Note: It is not necessary to mark in the fences when creating a boxplot.

Worked example 3

The *age* of 14 children on a cruise ship was recorded.

12 2 4 10 7 10 8 16 8 7 10 1 15 9

a. Calculate the lower and upper fences of the data set.

Explanation

Step 1: Arrange the data in ascending order.

1 2 4 7 7 8 8 9 10 10 10 12 15 16

Step 2: Determine the lower half and upper half of the data set.

$n = 14$

Position of median: $\frac{14+1}{2} = 7.5$

Lower half: 1 2 4 7 7 8 8

Upper half: 9 10 10 10 12 15 16

Continues →

Step 3: Determine the value of Q_1 and Q_3.

Q_1 is the median of the lower half of the data set.

$Q_1 = 7$

Q_3 is the median of the upper half of the data set.

$Q_3 = 10$

Step 4: Calculate the IQR.

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 10 - 7 \\ &= 3 \end{aligned}$$

Step 5: Calculate the lower and upper fences.

$$\begin{aligned} lower\ fence &= Q_1 - (1.5 \times IQR) \\ &= 7 - (1.5 \times 3) \\ &= 2.5 \end{aligned}$$

$$\begin{aligned} upper\ fence &= Q_3 + (1.5 \times IQR) \\ &= 10 + (1.5 \times 3) \\ &= 14.5 \end{aligned}$$

Answer

Lower fence: 2.5 years old

Upper fence: 14.5 years old

b. Identify any outliers.

Explanation

1 and 2 are both less than the lower fence (2.5).

15 and 16 are both greater than the upper fence (14.5).

Answer

1, 2, 15, 16

Constructing and interpreting boxplots

See worked example 4

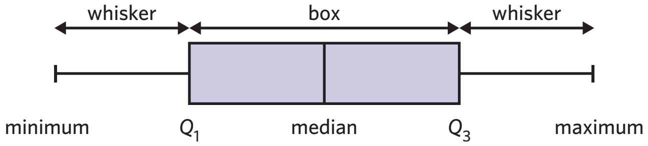
A **boxplot** is a graphical representation of a five-number summary as well as any outliers.

If there are no outliers, the leftmost and rightmost ends of the whiskers represent the minimum and maximum values. If there is an outlier, the whisker ends at the most extreme value that is not an outlier.

The left and right borders of the box represent Q_1 and Q_3 respectively.

The vertical line in the centre of the box represents the median.

Outliers are indicated by dots which lie outside the range of the box and whiskers.



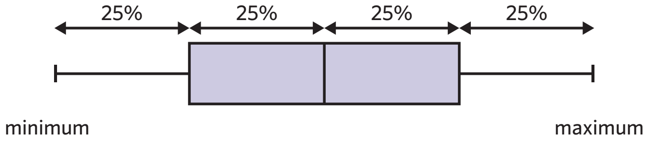
See worked example 5

See worked example 6

Boxplots can display data sets with a large number of numerical values. A calculator can be helpful in creating a boxplot directly from the data.

The distance between two consecutive boundaries represents 25% of data. The following intervals all represent 25% of data.

- Minimum to Q_1
- Q_1 to median
- Median to Q_3
- Q_3 to maximum



Worked example 4

The five-number summary for the following stem plot is:

1, 7, 16.5, 23, 39

Key: 1 | 2 = 12

0	1	2	3	5	5	9
1	0	0	1	5	8	9
2	1	2	3	3	5	
3	3	7	9			

Construct a boxplot using the data.

Explanation

Step 1: Check for any outliers.

Calculate the IQR.

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 23 - 7 \\ &= 16 \end{aligned}$$

Calculate the lower fence.

$$\begin{aligned} lower\ fence &= Q_1 - (1.5 \times IQR) \\ &= 7 - (1.5 \times 16) \\ &= -17 \end{aligned}$$

Calculate the upper fence.

$$\begin{aligned} upper\ fence &= Q_3 + (1.5 \times IQR) \\ &= 23 + (1.5 \times 16) \\ &= 47 \end{aligned}$$

Locate any outliers.

All data values lie between the lower and upper fences. Therefore, there are no outliers.

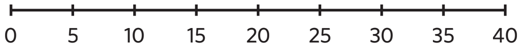
Step 2: Construct an axis with an appropriate scale.

The scale should cover a range such that it shows all data and outliers.

Intervals should be easy to work with.

Data ranges from 1 to 39, so an appropriate scale would run from 0 to 40.

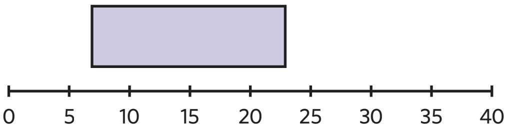
An interval of five makes the larger range easier to work with.



Step 3: Draw the border of the box.

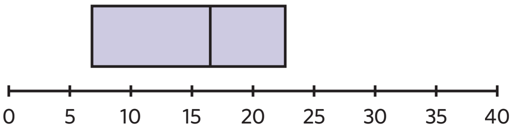
The value of the left border is Q_1 and the value of the right border is Q_3. As given by the five-number summary, Q_1 is 7 and Q_3 is 23.

The height of the box is not important.



Step 4: Mark in the vertical line.

The value of the vertical line in the middle of the box is the median. As given by the five-number summary, the median is 16.5.

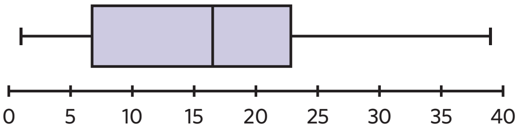


Step 5: Draw the whiskers.

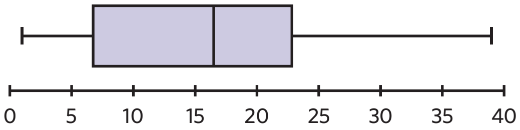
As there is no outlier, the left whisker ends at the minimum value and the right whisker ends at the maximum value.

Draw a whisker ranging from the minimum (1) to the leftmost point of the box.

Draw a whisker ranging from the rightmost point of the box to the maximum (39).



Answer



Worked example 5

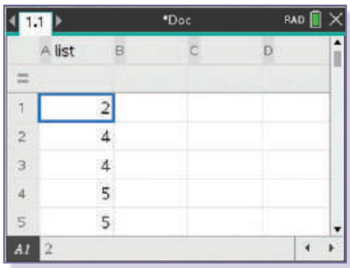
Construct a boxplot for the following data set.

2 4 4 5 5 5 7 6 5 4 6 5 6 7 8 12 3 5

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select ‘1: New’ → ‘4: Add Lists & Spreadsheet’.

Step 2: Name column A 'list' and enter the data values into column A, starting from row 1.



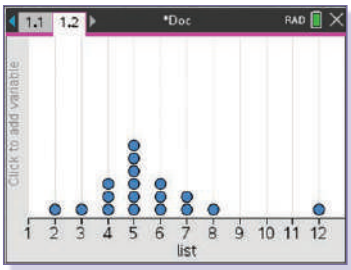
	A list	B	C	D
=				
1	2			
2	4			
3	4			
4	5			
5	5			

A1 2

Step 3: Press ctrl + doc▾ and select ‘5: Add Data & Statistics’.

Move the cursor to the horizontal axis and select ‘Click to add variable’.

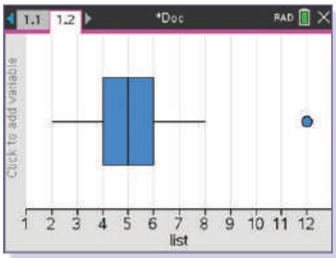
Select ‘list’.



Step 4: Press menu . Select ‘1: Plot Type’ → ‘2: Box Plot’.

Note: To change the view of the histogram press menu and use options within ‘5: Window/Zoom’.

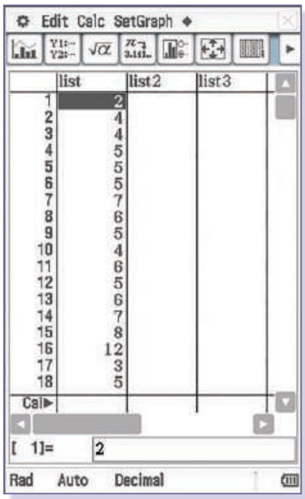
Answer



Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Rename list1 to 'list' and input the data values starting from row 1.



	list	list2	list3
1	2		
2	4		
3	4		
4	5		
5	5		
6	5		
7	7		
8	6		
9	5		
10	4		
11	6		
12	5		
13	6		
14	7		
15	8		
16	12		
17	3		
18	5		
Cal▸			

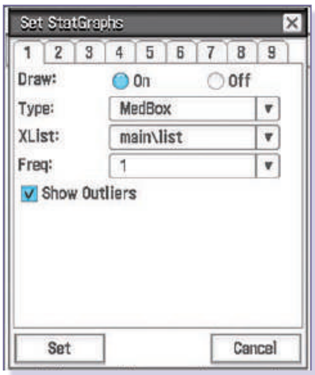
[1]= 2

Step 3: Configure the settings of the graph by tapping .

Create a histogram by changing ‘Type:’ to ‘MedBox’.

Specify the data set by changing ‘XList:’ to ‘main\list’.

Tick the ‘Show Outliers’ box.



Draw: ● On ○ Off

Type: MedBox

XList: main\list

Freq: 1

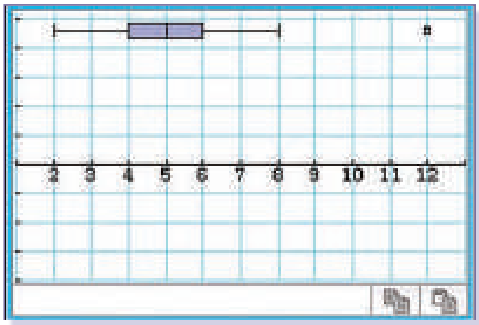
☑ Show Outliers

Tap ‘Set’ to confirm.

Step 4: Tap in the icon bar to plot the boxplot.

Continues →

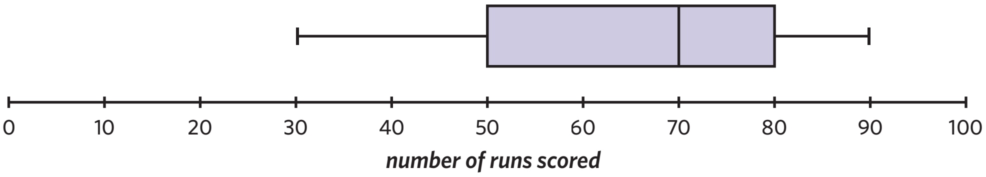
Answer



Worked example 6

The *number of runs scored* by the leading batsman at a local cricket club is represented by the following boxplot.

It is known that data was collected from 20 separate innings.



a. Determine the five-number summary for the data and identify any outliers.

Explanation

Step 1: Locate any potential outliers.

There are no potential outliers indicated on the boxplot.

Step 2: Locate and identify the minimum and maximum values.

As there are no outliers, the leftmost and rightmost ends of the whiskers represent the minimum and maximum values respectively.

$minimum = 30$

$maximum = 90$

Step 3: Locate and identify the value of Q_1 and Q_3.

The left and right borders of the box represent Q_1 and Q_3 respectively.

$Q_1 = 50$

$Q_3 = 80$

Step 4: Locate and identify the median.

The vertical line in the centre of the box represents the median.

$median = 70$

Answer

30, 50, 70, 80, 90

There are no outliers.

b. Between which two values does the middle 50% of data lie?

Explanation

The middle 50% of data lies between Q_1 and Q_3.

$Q_1 = 50$

$Q_3 = 80$

Answer

50 and 80 runs

Continues →

c. In approximately what percentage of innings did they score more than 80 runs?

Explanation

$Q_3 = 80$

25% of data is greater than Q_3.

Answer

25%

d. In approximately how many innings did they score at least 50 runs, but fall short of 100?

Explanation

Step 1: Identify the percentage of data that lies above 50 runs but less than 100 runs.

$Q_1 = 50$

$maximum = 90$

75% of data lies between Q_1 and the maximum.

Step 2: Calculate the number of innings.

There were 20 separate innings.

Calculate 75% of 20:

$0.75 \times 20 = 15$

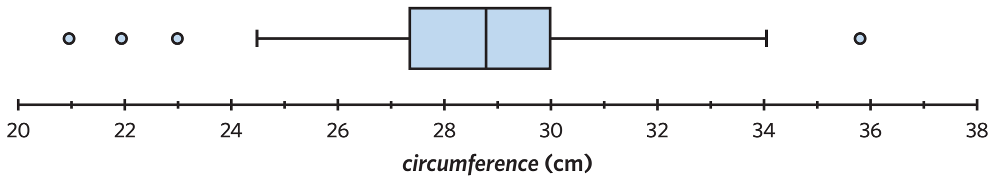
Answer

15 innings

Exam question breakdown

VCAA 2017 Exam 1 Data analysis Q2

The boxplot shows the distribution of the forearm *circumference*, in centimetres, of 252 people.



The five-number summary for the forearm *circumference* of these 252 people is closest to

A. 21, 27.4, 28.7, 30, 34

B. 21, 27.4, 28.7, 30, 35.9

C. 24.5, 27.4, 28.7, 30, 34

D. 24.5, 27.4, 28.7, 30, 35.9

E. 24.5, 27.4, 28.7, 30, 36

Explanation

Locate and identify the minimum and maximum values.

As there are outliers, the leftmost and rightmost dots represent the minimum and maximum values respectively.

$minimum = 21$

$maximum = 35.9$

The only option with a minimum of 21 and maximum of 35.9 is option B.

Answer

B

58% of students answered this question correctly.

38% of students incorrectly answered option C. Students who answered C ignored the outlier points when determining the maximum and minimum values for the five-number summary. Even though the points are identified as outliers, they are still valid data points within the data set and must be used as maximum and minimum values if appropriate.

1E Questions

Calculating the five-number summary

1. Find the value of Q_3 for the following data set.

7 9 3 8 4 10 2 6 6 4

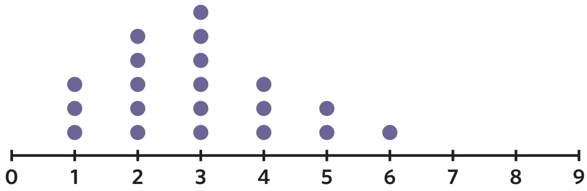
A. 4 **B.** 6 **C.** 8 **D.** 10

2. A physics class made paper aeroplanes to test aerodynamics. The *distance* (m) that each paper aeroplane flew was recorded.

34 26 27 13 18 16 6 11 25 23 33 31 9 28 20 17 32 35 9 35

Construct a five-number summary for the data.

3. Construct a five-number summary for the following dot plot.



4. Construct a five-number summary for the following stem plot.

Key: 1 | 7 = 17

0	7 7 8
1	2 3 5 7 9
2	1 1 3 5
3	1 2
4	1

Calculating the range and interquartile range

5. Which of the following statements is true?

A. The IQR is always greater than the range.

B. The IQR is always greater than or equal to the range.

C. The IQR is always less than the range.

D. The IQR is always less than or equal to the range.

6. The following data shows the amount of money, correct to the nearest dollar, that Alice spent at her school canteen each day over 2 weeks.

1 4 0 3 6 7 4 0 4 2

a. Find the range of the amount of money Alice spent.

b. Find the IQR.

7. At the annual cheese rolling festival, the *finishing time*, in seconds, of each cheese is recorded in the following stem plot.

Key: 5 | 1 = 5.1 secs

5	1 6 6 9
6	2 2 3 7 9
7	2 8
8	1

What is the IQR?

Identifying outliers

8. 100 Year 12 students were asked how much money they have spent on games in the past year.

The five-number summary for the results is: 0, 50, 90, 230, 395.

The lower fence for the data is

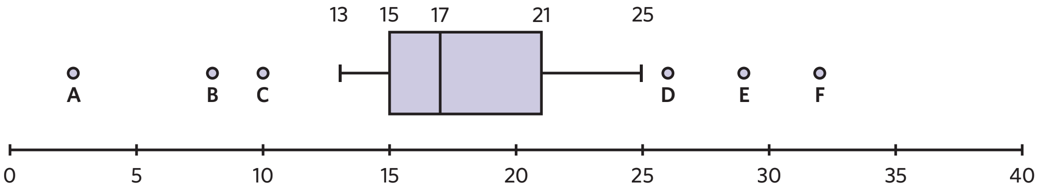
A. −$220

B. −$130

C. $0

D. $50

9. Identify which data values have been incorrectly marked as outliers on the boxplot.



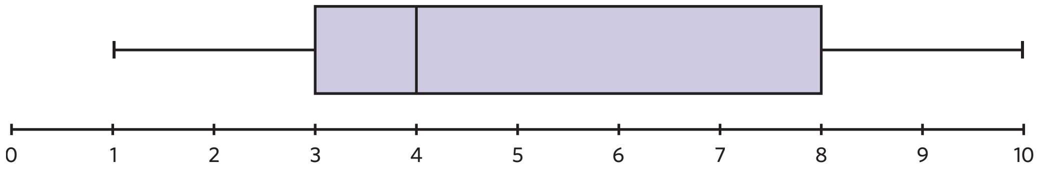
10. Identify any outliers in the following data sets.

a. 3 7 7 8 8 9 9 9 10 14 15

b. 25 24 28 20 30 24 25 24 26 26 19 24

Constructing and interpreting boxplots

11.



The five-number summary for the data shown in the boxplot is

A. 1, 3, 4, 8, 10

B. 3, 3, 4, 8, 8

C. 1, 4, 4, 4, 10

D. 2, 3, 4, 8, 9

12. A group of 20 students, each from different schools, competed in an interschool diving competition. They each performed one dive and received a *score* from 1 to 100.

62 100 39 50 93 25 53 53 44 43 40 28 36 30 84 84 59 100 58 76

a. Use a calculator to construct a boxplot from the data.

b. The percentage of students that received a *score* of more than 80 is closest to

A. 25%

B. 50%

C. 75%

D. 100%

c. The percentage of students that received a *score* between 39.5 and 80 is closest to

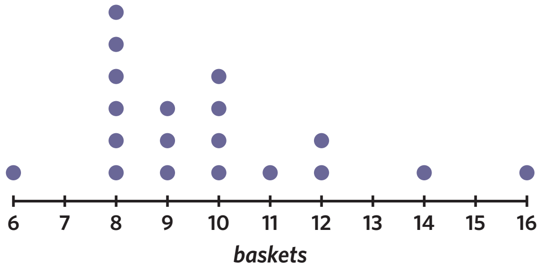
A. 25%

B. 50%

C. 75%

D. 100%

13. A team of basketball players held a mini tournament to see how many *baskets* they can shoot in one minute. The results are shown in the dot plot.



The five-number summary is: 6, 8, 9, 11, 16.

a. Construct a boxplot by hand to represent the data in the dot plot.

b. The percentage of players that scored more than 9 *baskets* is closest to

A. 25%

B. 50%

C. 75%

D. 100%

c. The percentage of players that scored 11 or less *baskets* is closest to

A. 25%

B. 50%

C. 75%

D. 100%

Joining it all together

14. The *goal tally* of the top 20 forwards in the 2000 AFL season is recorded.

51 47 52 70 57 53 56 76 42 53 50 109 49 43 47 44 42 69 40 68

a. Construct a five-number summary for the data.

b. Calculate the IQR.

c. Identify any outliers.

d. Construct a boxplot by hand to represent the data.

15. The owner of a record store documents the number of *records sold* on 17 different days.

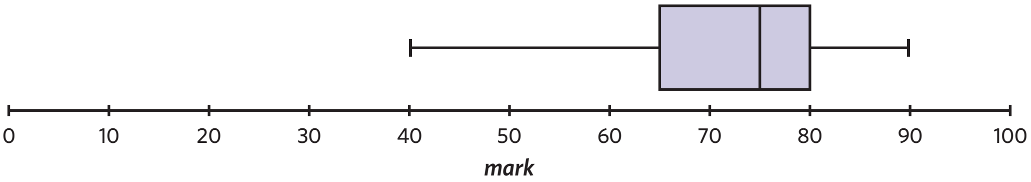
12 11 7 15 8 10 11 13 24 7 16 13 11 15 8 12 7

a. Use a calculator to construct a boxplot.

b. Identify any outliers.

c. Construct a five-number summary for the data.

16. The *mark*, out of 100, for 20 students who sat a university entrance exam was recorded.



a. Use the boxplot to estimate the percentage of students that scored between:

i. 40 and 65

ii. 65 and 80

iii. 75 and 90

iv. 65 and 90

b. The data is correct, but the actual boxplot has been drawn incorrectly. Identify the error and redraw the boxplot given that the second lowest score was 50, the second highest score was 85 and Q_1, the median, and Q_3 are correctly positioned.

Exam practice

17. The following stem plot shows the distribution of mathematics *test scores* for a class of 23 students.

Key: 4 | 2 = 42 n = 23

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

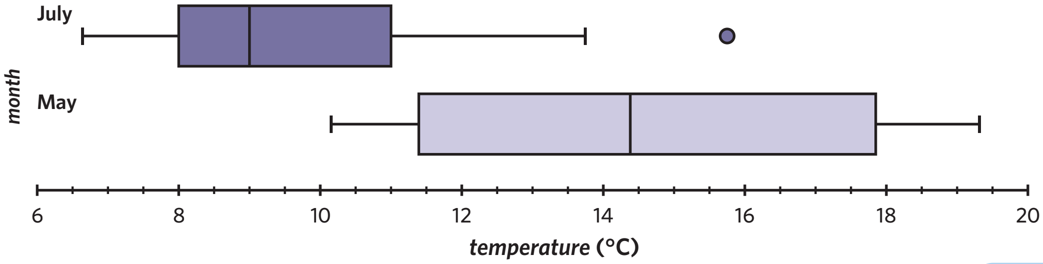
For this class, the interquartile range (IQR) of *test scores* is

A. 14.5 **B.** 17.5 **C.** 18
D. 24 **E.** 49

VCAA 2019 Exam 1 Data analysis Q5

87% of students answered this question correctly.

18. The following boxplots display the distribution of maximum daily *temperature* for the months of May and July.

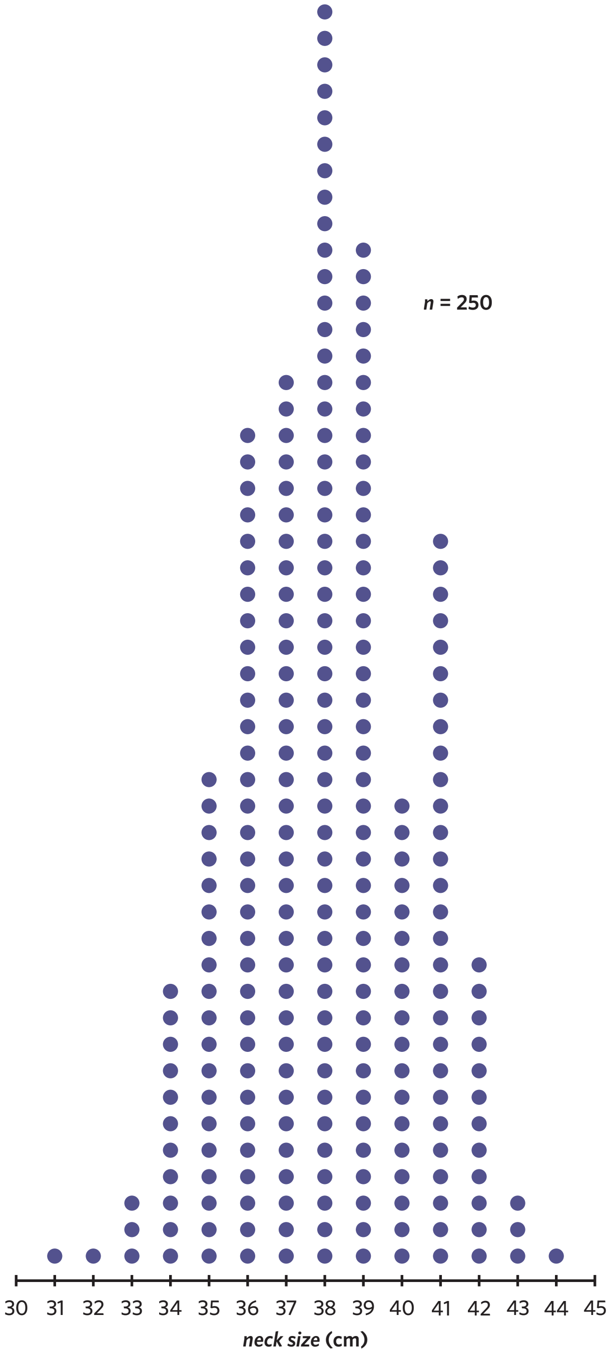


Determine the value of the upper fence for the July boxplot. (1 MARK)

VCAA 2016 Exam 2 Data analysis Q2bii

60% of students answered this question correctly.

19. The *neck size*, in centimetres, of 250 men was recorded and displayed in the following dot plot.



The five-number summary for this sample of neck sizes, in centimetres, is given in the following table.

minimum	first quartile (Q_1)	median	third quartile (Q_3)	maximum
31	36	38	39	44

Use the five-number summary to construct a boxplot, showing any outliers if appropriate. (2 MARKS)

VCAA 2020 Exam 2 Data analysis Q2c

The average mark on this question was **1.1**.

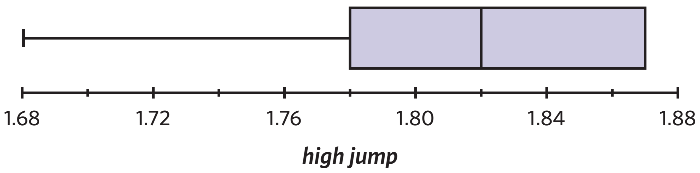
20. In the sport of heptathlon, athletes compete in seven events.

These events are the 100 m hurdles, high jump, shot-put, javelin, 200 m run, 800 m run and long jump.

Fifteen female athletes competed to qualify for the heptathlon at the Olympic Games.

Their results for three of the heptathlon events – high jump, shot-put and javelin – are shown in the table.

The following boxplot was constructed to show the distribution of high jump heights for all 15 athletes in the qualifying competition.



Explain why the boxplot has no whisker at its upper end. (1 MARK)

VCAA 2021 Exam 2 Data analysis Q1e

athlete number	*high jump* (metres)	*shot-put* (metres)	*javelin* (metres)
1	1.76	15.34	41.22
2	1.79	16.96	42.41
3	1.83	13.87	46.53
4	1.82	14.23	40.53
5	1.87	13.78	40.62
6	1.73	14.50	45.62
7	1.68	15.08	42.33
8	1.82	13.13	40.88
9	1.83	14.22	39.22
10	1.87	13.62	42.51
11	1.87	12.01	42.75
12	1.80	12.88	38.12
13	1.83	12.68	42.65
14	1.87	12.45	41.32
15	1.78	11.31	42.88

47% of students answered this question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

21. George works for a furniture company that has a number of stores across Victoria. He is asked to investigate the association between *number of couches sold* and *store location* (Shepparton, Fitzroy, Vermont, Bundoora). These variables are

A. a numerical variable and an ordinal variable respectively.

B. a numerical variable and a nominal variable respectively.

C. an ordinal variable and a numerical variable respectively.

D. both numerical variables.

E. both categorical variables.

Adapted from VCAA 2017NH Exam 1 Data analysis Q4

Recursion and financial modelling *Year 11 content*

22. Sheldon invests \$20 000 for a total of 365 days.

His interest compounds daily and he does not touch his investment for its entire duration.

If it reaches a value of \$20 506.28 at the end of the 365 days, the interest rate per annum is closest to

A. 1.0% p.a

B. 1.5% p.a.

C. 2.0% p.a.

D. 2.5% p.a.

E. 3.0% p.a.

Adapted from VCAA 2014 Exam 1 Business-related mathematics Q3

Data analysis *Year 11 content*

23. Students in a biology lab are experimenting to see the relationship between the number of eggs in a termite colony and its size by the end of the trial period.

The line of good fit for this data is:

size of colony = −4 + 1.05 × *number of eggs*

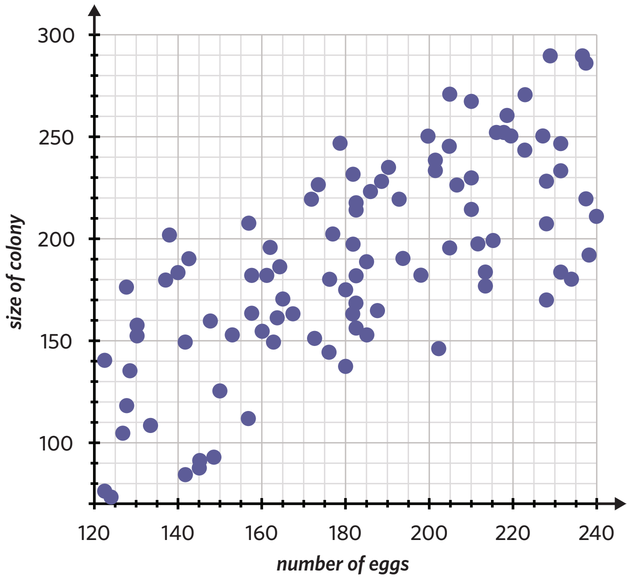
a. Draw the line of good fit on the scatterplot. (1 MARK)

b. Suppose the students had started one group with exactly 100 eggs. Estimate the population projection for this group of experiments.

Round your answer to the nearest whole number. (1 MARK)

c. In reference to the prediction made in part **b**, is this an example of interpolation or extrapolation? (1 MARK)

Adapted from VCAA 2018NH Exam 1 Data analysis Q5



1F Describing numerical data

STUDY DESIGN DOT POINTS

- representation, display and description of the distributions of numerical variables: dot plots, stem plots, histograms; the use of a logarithmic (base 10) scale to display data ranging over several orders of magnitude and their interpretation in terms of powers of ten
- use of the distribution(s) of one or more categorical or numerical variables to answer statistical questions

1A 1B 1C 1D 1E 1F 1G 1H 1I

KEY SKILLS

During this lesson, you will be:

- describing the distribution of histograms
- describing the distribution of dot plots and stem plots
- identifying the best measure of centre
- describing the distribution of boxplots.

KEY TERMS

- Positively skewed
- Negatively skewed
- Symmetric
- Bimodal
- Centre

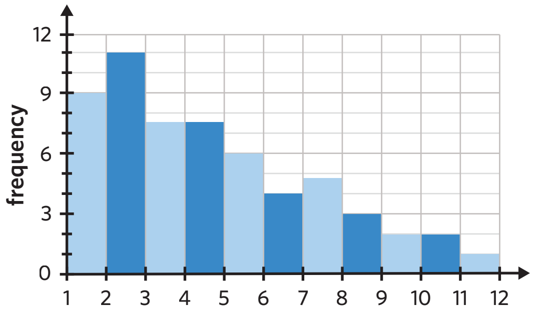
Once numerical data has been collected and graphed, it is important to be able to describe what it displays. This allows people using the data to understand exactly what the data is suggesting, and helps to convey underlying trends.

Describing the distribution of histograms

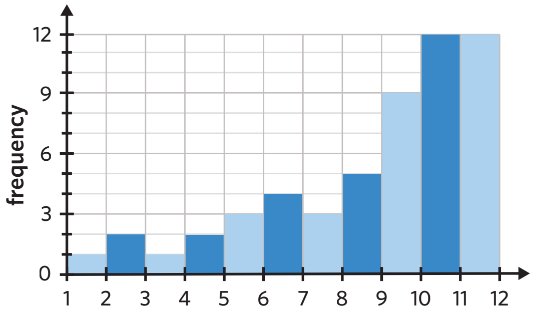
Numerical data distributions are generally described in terms of their shape, centre and spread.

Shape

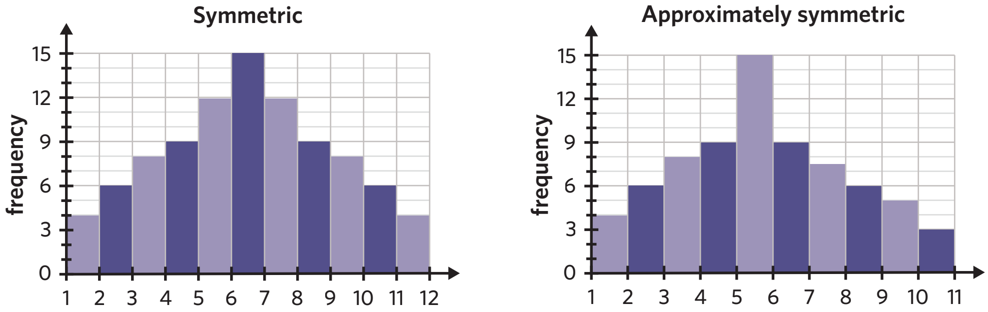
A **positively skewed** distribution trails off in a positive direction on the horizontal axis.



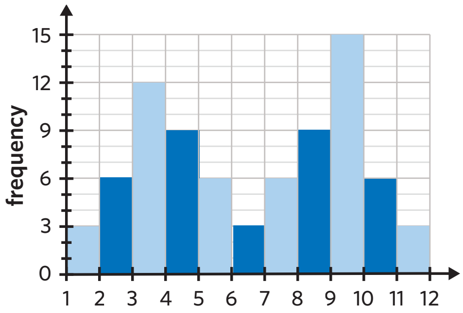
A **negatively skewed** distribution trails off in a negative direction on the horizontal axis.



A **symmetric** distribution is the same on both sides of the centre. If the distribution isn't exactly symmetric, it is important to describe the shape as approximately symmetric.

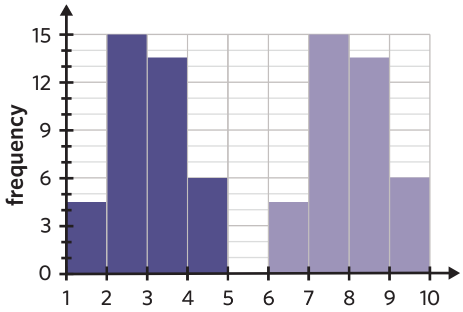


A symmetric or approximately symmetric distribution can also be **bimodal**. This occurs when there are two distinct peaks in the distribution. These peaks do not necessarily have to be equal.



Centre

The **centre** refers to the middle of a distribution. The following two distributions have the exact same shape, but different centres.



Either the mean or median can be used as a measure of centre. This lesson will focus on the median.

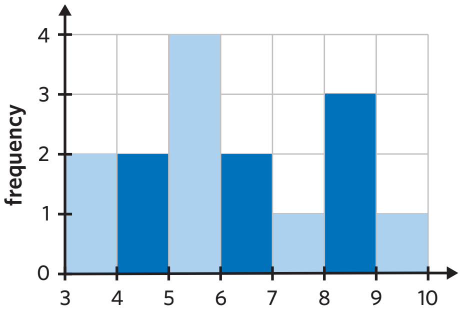
Spread

Recall that the spread refers to the spacing of values within a data set.
For histograms and other numerical distributions, the range can be used as the measure of spread.

$range = maximum\ value - minimum\ value$

This histogram has a range of 7.

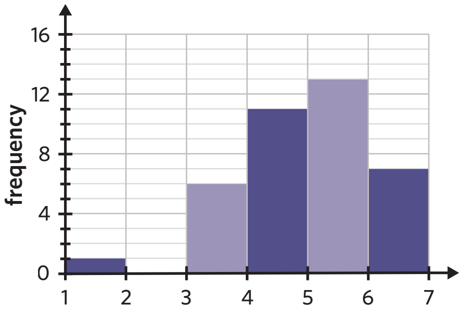
$range = 10 - 3 = 7$



It is also important to identify potential outliers in a histogram. These are values that fall outside of what looks normal or reasonable, and can be identified by eye. This means they need to be referred to as potential outliers.

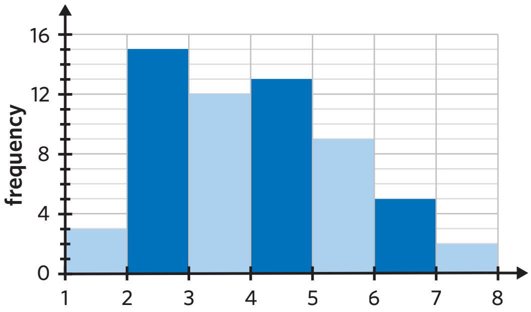
This data set has a potential outlier in the interval 1–<2.

Note: The interquartile range (IQR) is often used as a more reliable measure of spread when outliers are present. This lesson however only focuses on range as the measure of spread.



Worked example 1

Consider the histogram.



a. Estimate the median.

Explanation

Step 1: Calculate the number of data points by summing the frequency of each column.

$3 + 15 + 12 + 13 + 9 + 5 + 2 = 59$

Step 2: Determine the position of the median.

Remember that the median is located in the $\left(\frac{n+1}{2}\right)^{\text{th}}$ position.

In this case, $n = 59$.

$$\frac{59+1}{2} = \frac{60}{2}$$
$$= 30$$

Step 3: Determine in which interval the 30th data point lies by calculating the cumulative frequency for each interval.

interval	cumulative frequency
1–<2	3
2–<3	3 + 15 = 18
3–<4	18 + 12 = 30

The 30th data point occurs in the 3–<4 interval.

Step 4: Estimate the median.

$$\frac{3+4}{2} = \frac{7}{2}$$
$$= 3.5$$

Answer

Approximately 3.5

b. Describe the histogram in terms of shape, spread and potential outliers.

Explanation

Step 1: Determine the shape of the histogram.

The majority of the data is towards the left of the distribution.

The distribution trails off in the positive direction.

This means the histogram is positively skewed.

Step 2: Calculate the spread of the histogram.

$$\textit{range} = \textit{maximum value} - \textit{minimum value}$$
$$= 8 - 1$$
$$= 7$$

Continues →

Step 3: Identify any potential outliers.

There are no data values that fall outside of what looks reasonable.

Answer

The histogram is positively skewed with a range of 7 and no potential outliers.

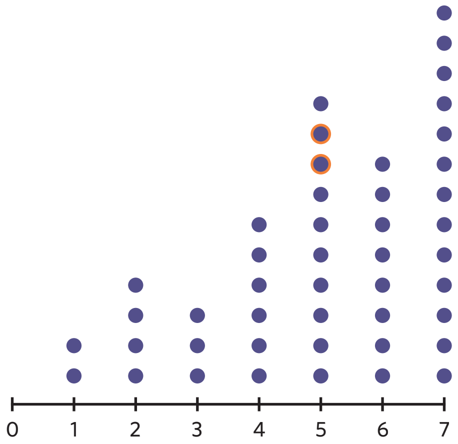
Describing the distribution of dot plots and stem plots

Dot plots and stem plots are also described using shape, centre and spread.

When calculating the median for dot plots and stem plots, the exact location of the median can be identified, unlike in histograms where only an approximate value can be calculated.

For example, consider the following dot plot.

See worked example 2



The median is the average of the 23rd and 24th values. These are highlighted on the dot plot.

The median is $\frac{5 + 5}{2} = 5$.

Stem plots are displayed vertically, not horizontally like histograms and dot plots. The distribution will still trail off in the same direction (in the positive or negative direction) for a positively or negatively skewed distribution.

See worked example 3

Positively skewed

Key: 4 | 0 = 40

4	0	1	4	6	7	7	8
5	3	4	4	5	6		
6	0	1	1	3	9		
7	1	4	5	8			
8	2	6					
9	1						

Negatively skewed

Key: 4 | 0 = 40

4	0						
5	3	4	4				
6	0	1	1				
7	1	4	5	8			
8	2	3	3	5	7	9	
9	1	2	4	4	6	7	8

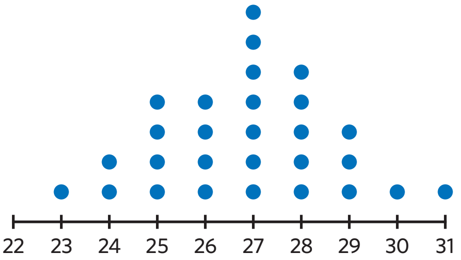
Approximately symmetric

Key: 4 | 0 = 40

4	0					
5	3	4	4			
6	0	1	1	3	9	
7	1	4	5	8	8	9
8	2	6				
9	1					

Worked example 2

Consider the dot plot.



a. Determine the median.

Explanation

Step 1: Count the number of data points.

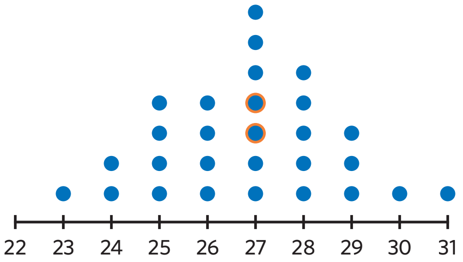
There are 28 data points.

Step 2: Determine the location of the median.

$$\frac{28 + 1}{2} = \frac{29}{2}$$
$$= 14.5$$

The median will be the average of the 14th and 15th data points.

Step 3: Locate the data points and calculate the median.



$$\frac{27 + 27}{2} = 27$$

Answer

27

b. Describe the dot plot in terms of shape, spread and potential outliers.

Explanation

Step 1: Determine the shape of the dot plot.

The majority of the data is in the middle of the distribution.

The distribution trails off in both directions approximately equally.

This means the dot plot is approximately symmetric.

Step 2: Calculate the spread of the dot plot.

$$\begin{aligned} range &= maximum\ value - minimum\ value \\ &= 31 - 23 \\ &= 8 \end{aligned}$$

Step 3: Identify any potential outliers.

There are no data values that fall outside of what looks reasonable.

Answer

The dot plot is approximately symmetric with a range of 8 and no potential outliers.

Worked example 3

Consider the stem plot.

Key: 40 | 7 = 407

40	7
41	
42	
43	
44	4 6
45	1 3 3
46	2 6 7
47	4 7 8 9
48	0 0 4 5 6 8
49	1 2 5 5 5 9
50	3 4 4 7 8 8 9 9

a. Calculate the median.

Explanation

Step 1: Count the number of data points.

There are 33 data points.

Step 2: Determine the location of the median.

$$\frac{33 + 1}{2} = \frac{34}{2} = 17$$

The median will be the 17th data point.

Step 3: Locate the data point.

Key: 40 | 7 = 407

40	7
41	
42	
43	
44	4 6
45	1 3 3
46	2 6 7
47	4 7 8 9
48	0 0 4 5 6 8
49	1 2 5 5 5 9
50	3 4 4 7 8 8 9 9

Answer

485

b. Describe the stem plot in terms of shape, spread and potential outliers.

Explanation

Step 1: Determine the shape of the stem plot.

The majority of the data is towards the larger values in the distribution.

The distribution trails off in the negative direction.

This means the stem plot is negatively skewed.

Step 2: Calculate the spread of the stem plot.

$$\begin{aligned} \textit{range} &= \textit{maximum value} - \textit{minimum value} \\ &= 509 - 407 \\ &= 102 \end{aligned}$$

Step 3: Identify any potential outliers.

There is one data point (407) that appears to lie outside the normal range of data.

Continues →

Answer

The stem plot is negatively skewed with a range of 102 and one potential outlier.

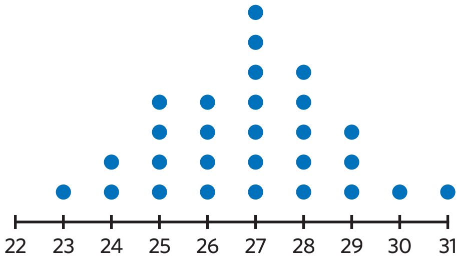
Identifying the best measure of centre

Either the mean or the median can be used as the best measure of centre for a distribution of data.

The mean is the best measure of centre for distributions that have no outliers and are approximately symmetric.

For skewed distributions or distributions with outliers, the median is the best measure of centre as it refers to the middle value and doesn't include the numerical value of any outliers.

Consider the dot plot. The median is 27, while the mean is 26.86.



The distribution is approximately symmetric with no outliers. Therefore, the mean hasn't been affected by any skewed data, and provides a more precise measure of centre than the median.

Consider the stem plot.

Key: 5 | 1 = 5.1

0	2							
1								
2								
3								
4								
5	1							
6	6	6						
7	2	5						
8	1	6	7	8				
9	4	5	5	6	8	9		
10	2	4	5	8	8	8	9	9

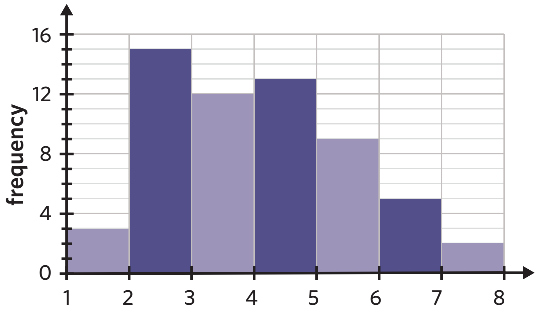
The median is 9.5, while the mean is 8.77. The mean is affected by the value of the outlier (0.2), so the median is the preferred measure of centre.

If the outlier is ignored, the median remains at 9.5 and the mean changes to 9.14. The median is still a better measure of centre because it is weighted more heavily to the bulk of the distribution.

Worked example 4

Identify the best measure of centre for the following distributions.

a.



Continues →

Explanation

Step 1: Identify the shape of the distribution.

The bulk of the data is towards the smaller values in the distribution.

The distribution trails off in the positive direction.

This means the histogram is positively skewed.

Step 2: Determine the best measure of centre.

When data is skewed, the best measure of centre is the median.

Answer

Median

b. **Key:** 4 | 0 = 40

4	0
5	3 4 4
6	0 1 1 3 9
7	1 4 5 8 8 9
8	2 6
9	1

Explanation

Step 1: Identify the shape of the distribution.

The majority of data is in the middle of the distribution.

The distribution trails off in both directions approximately equally.

This means the stem plot is approximately symmetric.

Step 2: Identify any potential outliers.

There are no data values that fall outside of what looks reasonable

Step 3: Determine the best measure of centre.

When data is approximately symmetric with no outliers, the best measure of centre is the mean.

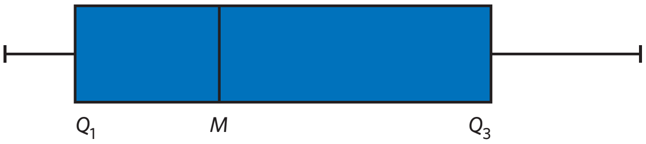
Answer

Mean

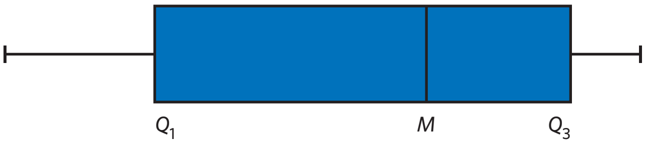
Describing the distribution of boxplots

Boxplots can also be described in terms of their shape, centre and spread. The same principles apply as with histograms, dot plots and stem plots however the method of interpreting the shape is slightly different.

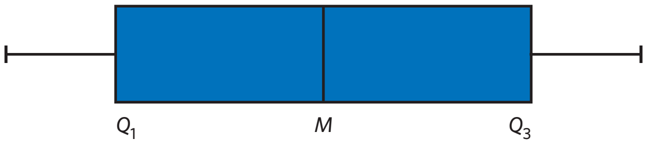
A positively skewed distribution trails off in the positive direction. For a boxplot, this means the median is located towards the left of the box. The left whisker will be short whereas the right whisker will be longer.



A negatively skewed distribution trails off in the negative direction. For a boxplot, this means the median is located towards the right of the box. The right whisker will be short whereas the left whisker will be longer.



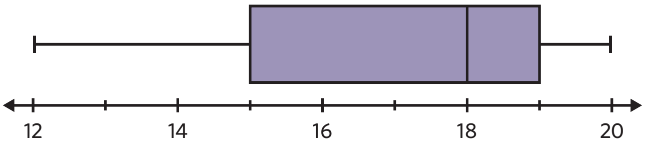
A symmetric distribution is evenly distributed and does not trail off in either direction more than the other. For a boxplot, this means the median is approximately in the centre of the box. The distances from the median to Q_1 and Q_3 are approximately equal, and the whiskers are also approximately equal.



Worked example 5

Consider the boxplot.

Describe the distribution in terms of shape, centre, spread and outliers.



Explanation

Step 1: Determine the shape of the boxplot.

The median is located towards the right of the box.

The distribution trails off in the negative direction.

This means the boxplot is negatively skewed.

Step 2: Identify the median of the boxplot.

The median is represented by the vertical line inside the box.

$median = 18$

Step 3: Calculate the spread of the boxplot.

$$\begin{aligned} range &= maximum\ value - minimum\ value \\ &= 20 - 12 \\ &= 8 \end{aligned}$$

Step 4: Identify any outliers.

There are no outliers in the boxplot.

Answer

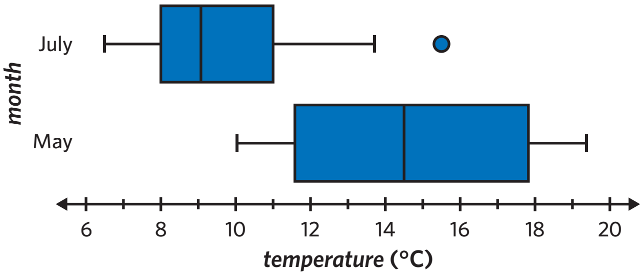
The boxplot is negatively skewed with no outliers, a median of 18 and a range of 8.

Exam question breakdown

VCAA 2016 Exam 2 Data analysis Q2bi

The following boxplots display the distribution of maximum daily *temperature* for the months of May and July.

Describe the shapes of the distributions of daily *temperature* (including outliers) for July and for May. (1 MARK)



Explanation

Step 1: Identify the shape of the distribution for July.

The median is located towards the left of the boxplot.

The distribution trails off in the positive direction.

The boxplot is positively skewed.

Step 2: Identify any outliers.

There is one outlier for July.

Step 3: Identify the shape of the distribution for May.

The median is located approximately in the middle of the boxplot.

Q_1 and Q_3 are located approximately the same distance from the median.

The whiskers are approximately the same length.

Step 4: Identify any outliers.

There are no outliers for May.

Continues →

Answer

July: Positively skewed with one outlier

May: Approximately symmetric with no outliers

56% of students answered this question correctly.

For this question, many students answered either symmetrically skewed, evenly distributed, bell-shaped or normally distributed for July. These are either incorrect descriptions of the shape of a distribution or inaccurate descriptions of the distribution for July.

1F Questions

Describing the distribution of histograms

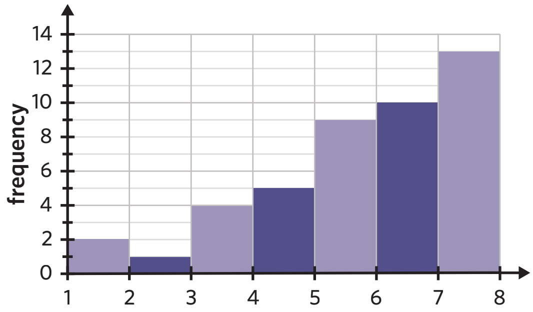
1. Describe the shape of the following histogram.

A. Positively skewed

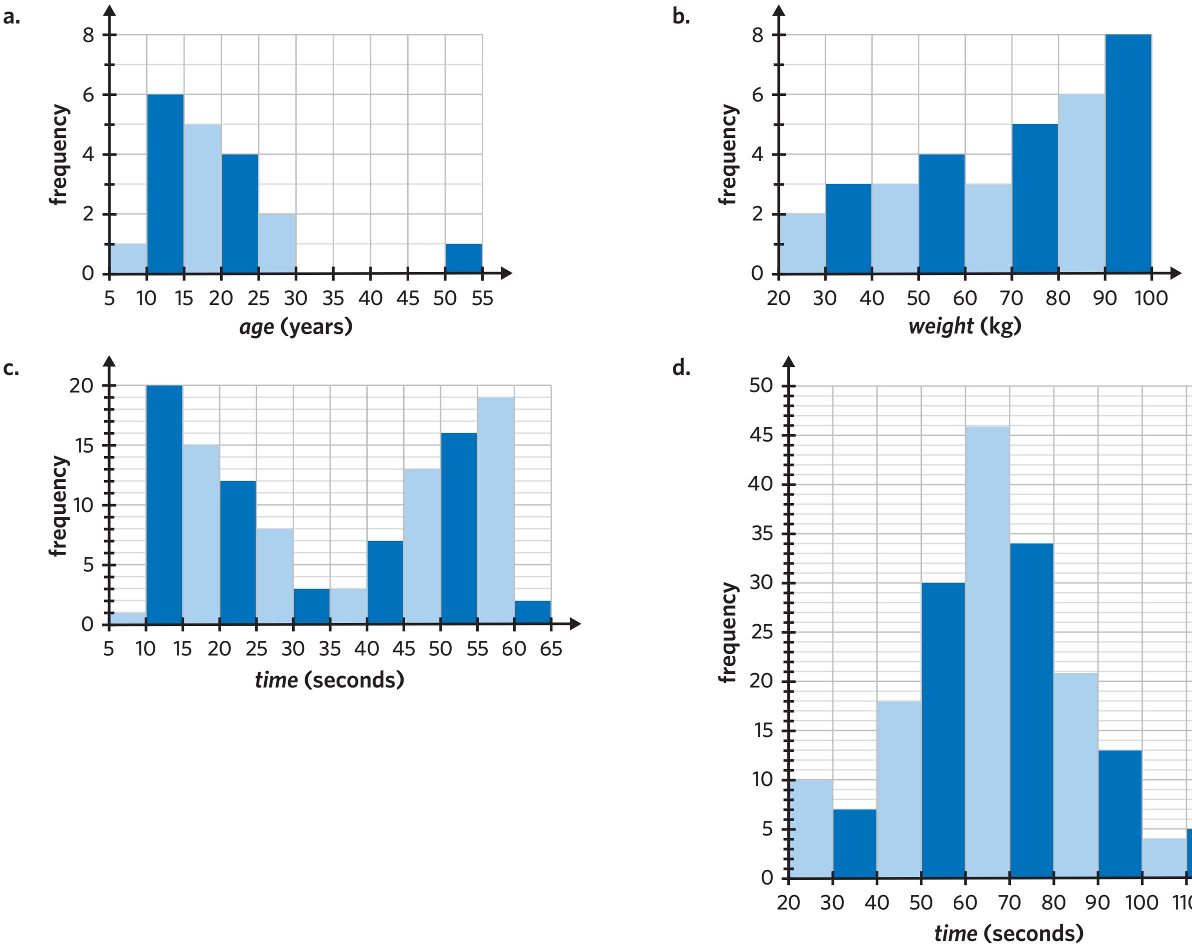
B. Positively skewed with a potential outlier

C. Negatively skewed

D. Approximately symmetric



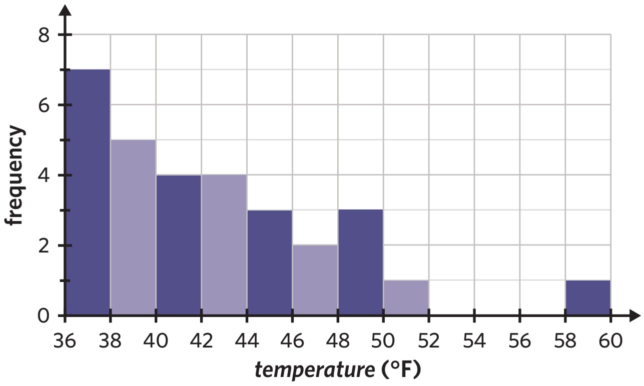
2. Describe the shape of each of the following histograms and identify any potential outliers.



3. The *temperature*, in °F, in New York was measured every day for the 30 days of November. The data has been placed into the following histogram.

a. Estimate the median *temperature*.

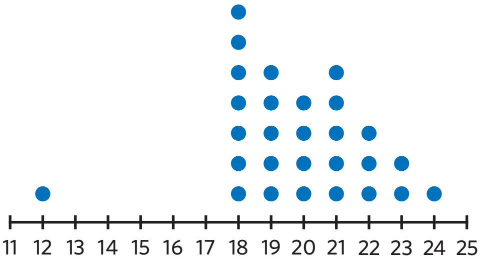
b. Describe the histogram in terms of shape, spread and potential outliers.



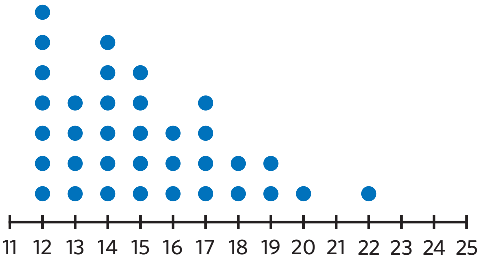
Describing the distribution of dot plots and stem plots

4. Which of the following dot plots is negatively skewed with a potential outlier?

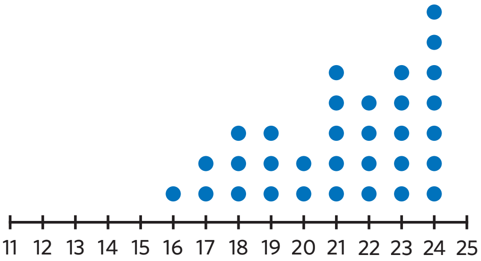
A.



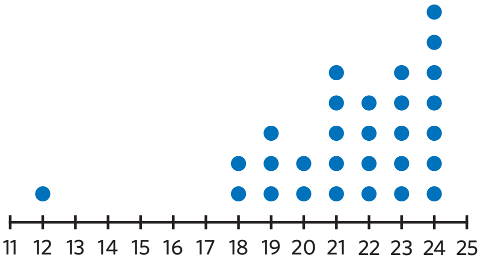
B.



C.

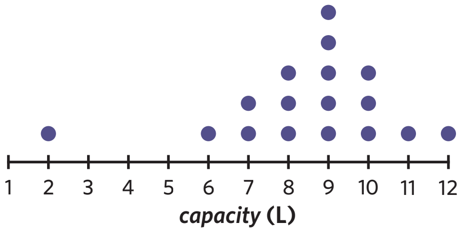


D.

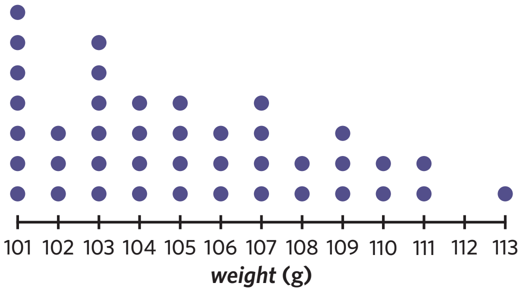


5. Describe the shape of each of the following dot plots and identify any potential outliers.

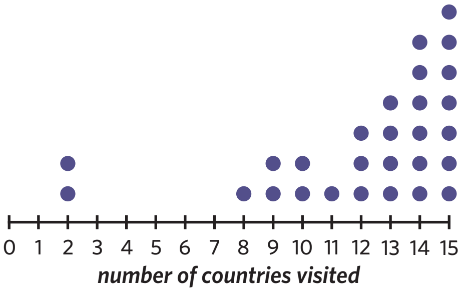
a.



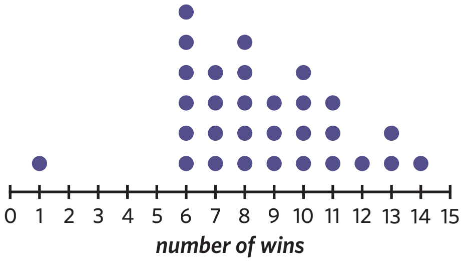
b.



c.



d.



6. Describe the shape of each of the following stem plots and identify any possible outliers.

a. **Key:** 1 | 7 = 17 cm

1	7
2	1 8
3	0 5 5
4	4 7
5	1 3 8
6	2 2 9
7	3 4 4 8
8	1 1 4 5 6 9

b. **Key:** 20 | 1 = 20.1 seconds

20	1
21	2 4 5
22	0 7 7
23	4 5 7 8 8
24	4 6 6 7 9 9
25	1 3 7 8
26	2 6
27	4 9
28	
29	
30	
31	9

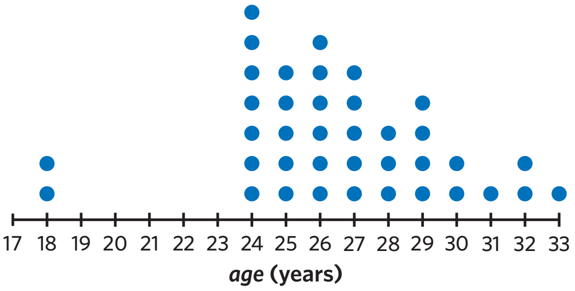
c. **Key:** 9 | 0 = 90 days

9	0 2 5 5 5 6 7 9
10	2 4 6 7 9 9
11	0 1 5
12	6 6 7 8
13	1 2 2
14	3
15	5
16	
17	
18	
19	4

d. **Key:** 200 | 4 = $2004

200	4
201	2 6
202	4 4 5 7
203	3 6 6 7 8
204	2 2 2 6 7
205	0 0 3 4 8 8 9
206	
207	
208	
209	3 5

7. Consider the following dot plot.



a. Calculate the median *age*.

b. Describe the dot plot in terms of shape, spread and potential outliers.

8. Sophie is a newspaper editor and records the *number of edits* she makes for each newspaper that gets printed. The results are shown in the following stem plot.

Key: 15 | 7 = 157

15	7
16	
17	
18	1 2
19	5 6 7 9
20	3 3 4 6 6 7
21	0 0 1 2 5 8 9 9
22	0 3 3 3 4 6 7 7 8 9

a. Calculate the median *number of edits*.

b. Describe the stem plot in terms of shape, spread and potential outliers.

Identifying the best measure of centre

9. Fill in the missing words.

The mean is the best measure of centre when the distribution is __________ with no __________.

A. negatively skewed, outliers

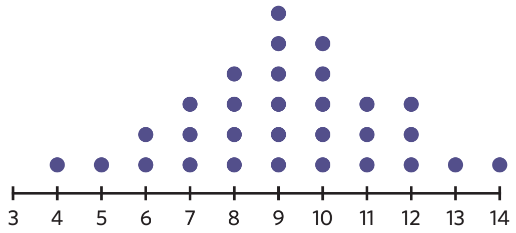
B. outliers, positively skewed

C. approximately symmetric, outliers

D. outliers, approximately symmetric

10. Identify whether the mean or median would be the best measure of centre for the following distributions.

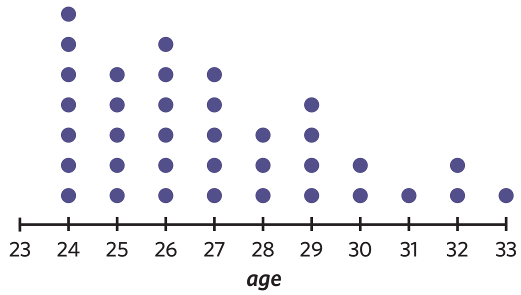
a.



b. **Key:** 0 | 3 = 3

0	3
1	
2	
3	
4	2 7
5	2 2 5
6	1 7 8
7	3 6 9 9
8	1 1 4 6 6 8
9	0 0 1 5 6 9
10	4 6 6 7 8 9 9 9

c.



d. **Key:** 50 | 2 = 50.2

50	2
51	3 4 6
52	3 8 8
53	0 3 6 8 8
54	2 2 6 7 8 9
55	0 1 8 9
56	2 6
57	5 8
58	
59	
60	
61	7

11. Rodney and Sally are discussing their answers to question **10d**. Rodney says the best measure of centre is the mean because the data is approximately symmetric. Sally says the best measure of centre is the median because there is a potential outlier. Who is correct and why?

Describing the distribution of boxplots

12. Fill in the blanks in the following statement.

For a negatively skewed boxplot, the median is towards the __________ of the __________.
The right whisker will be __________ whereas the left whisker will be __________.

A. left, centre, short, longer

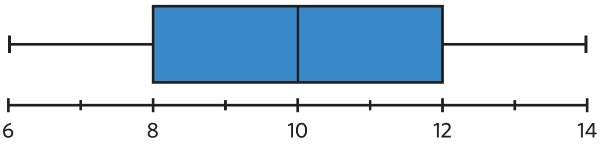
B. right, box, short, longer

C. middle, distribution, long, shorter

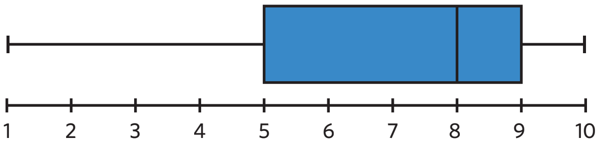
D. right, box, long, shorter

13. Describe the following boxplots in terms of shape, centre, spread and outliers.

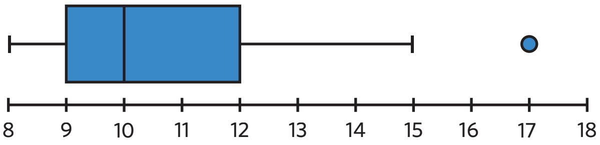
a.



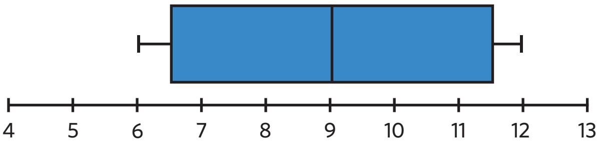
b.



c.



d.



Joining it all together

14. Which of the following statements is false?

A. A bimodal distribution has two clear peaks that are not necessarily equal.

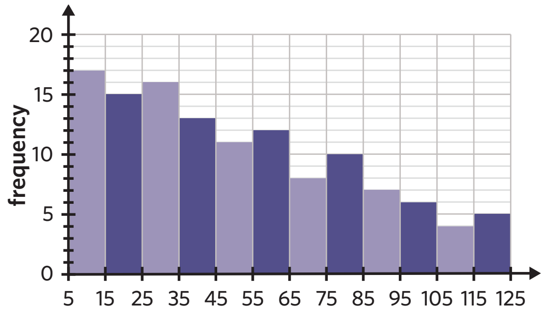
B. The median can only be estimated from a histogram but can be calculated exactly for both dot plots and stem plots.

C. The median should always be used as the best measure of centre.

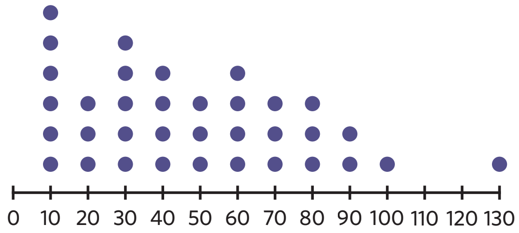
D. The median for a negatively skewed boxplot will be located towards the right side of the box.

15. Sashi constructed a distribution that is positively skewed with an outlier. The distribution has a median of 50 and a range of 120. Which of the following could be the distribution that Sashi constructed?

A.



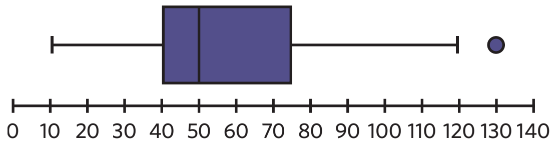
B.



C. **Key:** 1 | 8 = 18

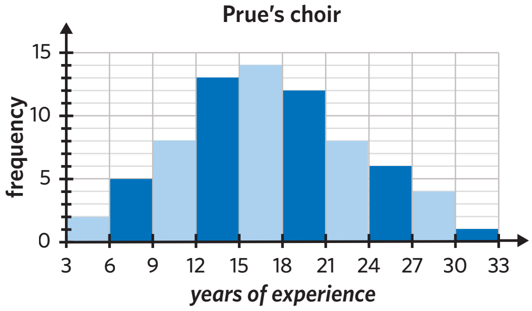
1	8
2	
3	
4	
5	5
6	3 4
7	0 1
8	5 7 8 8
9	1 1 6
10	3 4 8 9
11	0 0 0 5 8
12	2 5 6 6 6
13	1 2 5 7 7 8

D.



16. Prue and Gil are both singers, and are members of two different choirs. They each decided to collect data on the amount of singing experience each member of their choir has, in years. They produced the following distributions.

Prue's choir



Gil's choir

Key: 0 | 6 = 6 years

0	6 7
1	1 4 4
2	0 0 3 5 7 8 9
3	1 3 7 7
4	5
5	
6	
7	1

a. Describe both distributions in terms of shape, centre, spread and potential outliers.

b. Identify and explain the best measure of centre for each distribution.

Exam practice

17. The following histogram shows the distribution of *population size* of 48 countries in 2018.

The shape of this histogram is best described as

A. positively skewed with no outliers.

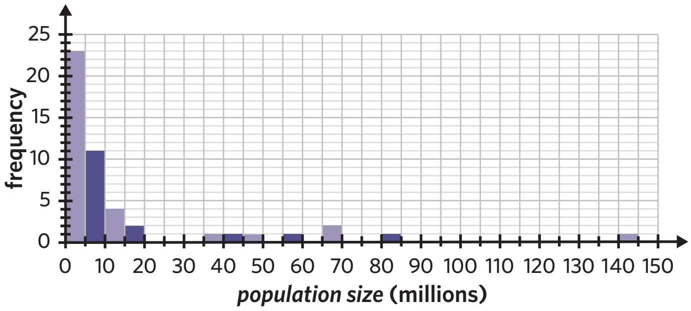
B. positively skewed with outliers.

C. approximately symmetric.

D. negatively skewed with no outliers.

E. negatively skewed with outliers.

VCAA 2019 Exam 1 Data analysis Q2



86% of students answered this question correctly.

18. Each dalmatian in a sample of 32 dalmatians had their *weight* recorded. The data is displayed in the following ordered stem plot.

Key: 21 | 6 = 21.6 kg $n = 32$

21	6 9 9
22	1 2 5 6
23	0 1 4 6 6 7 8
24	4 5 6 7 7 9
25	6 8
26	1 7 9
27	3 7
28	2
29	1 8
30	4
31	1

a. Describe the shape of the distribution. (1 MARK)

b. Determine the median *weight* for this group of dalmatians. (1 MARK)

Adapted from VCAA 2020 Exam 2 Data analysis Q1

Part **a**: **86%** of students answered this type of question correctly.

Part **b**: **70%** of students answered this type of question correctly.

19. The *times* between successive nerve impulses, in milliseconds, were recorded.

The following table shows the mean and the five-number summary calculated using 800 recorded data values.

The shape of the distribution of these 800 times is best described as

A. approximately symmetric.

B. positively skewed.

C. positively skewed with one or more outliers.

D. negatively skewed.

E. negatively skewed with one or more outliers.

VCAA 2020 Exam 1 Data analysis Q3

	***time* (milliseconds)**
mean	220
minimum value	10
first quartile (Q_1)	70
median	150
third quartile (Q_3)	300
maximum value	1380

Data: adapted from P Fatt and B Katz, 'Spontaneous subthreshold activity at motor nerve endings', The Journal of Physiology, 117, 1952, pp. 109-128

59% of students answered this question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

20. 25 students in a Year 11 class recorded their *heights* and displayed them in the following stem plot.

Key: 16 | 3 = 163 cm

16	3	3	4						
16	5	6	7	9	9				
17	0	0	0	0	1	1	1	3	4
17	5	7	7	7	9				
18	0	1	1						

The modal *height* is

A. 17 cm

B. 18 cm

C. 170 cm

D. 171 cm

E. 177 cm

Adapted from VCAA 2016 Exam 1 Data analysis Q3

Data analysis *Year 11 content*

21. A study examined the relationship between *IQ* and *test score* (%) on a logic test.

The following least squares regression equation was derived.

$\textit{test score} = -25.9 + 0.93 \times \textit{IQ}$

Which of the following conclusions drawn from the regression equation is true?

A. The correlation coefficient is 0.93.

B. A person's *test score*, as a percentage, can be determined by subtracting 25.9 from their *IQ*.

C. A person's *IQ* can be determined by subtracting 25.9 from their *test score*.

D. An increase of 1 point in *IQ* is associated with an increase of 0.93% in *test score*.

E. An increase of 1% in *test score* is associated with an increase of 0.93 points in *IQ*.

Adapted from VCAA 2018 Exam 1 Data analysis Q10

Recursion and financial modelling *Year 11 Content*

22. An electricity company, Edrolicity, charges interest for overdue bills. If a bill is not paid by the due date, interest is charged at a rate of 2.2% per month, compounding monthly.

a. Anna received an Edrolicity bill for $120. If she paid the bill one month after its due date, how much did she pay in total? (1 MARK)

b. Brody received a bill for $246, and failed to pay it by the due date. Write a recurrence relation in terms of T_0, T_{n+1} and T_n that shows the total amount of the bill n months after the due date. (2 MARKS)

c. Brody paid his bill five months after the due date. How much did Brody pay in interest only, correct to the nearest cent? (1 MARK)

Adapted from VCAA 2017 Exam 2 Recursion and financial modelling Q6

1G Introduction to standard deviation

STUDY DESIGN DOT POINT

- summary of the distributions of numerical variables; the five-number summary and boxplots (including the use of the lower fence ($Q_1 - 1.5 \times IQR$) and upper fence ($Q_3 + 1.5 \times IQR$) to identify and display possible outliers); the sample mean and standard deviation and their use in comparing data distributions in terms of centre and spread

1A 1B 1C 1D 1E 1F **1G** 1H 1I

KEY SKILLS

During this lesson, you will be:

- calculating the sample mean
- calculating the sample mean and standard deviation using technology.

KEY TERMS

- Population
- Sample
- Mean
- Standard deviation

The mean and standard deviation provide an important way for data analysts to investigate data sets. The mean gives a standard measure of centre, whilst the standard deviation is another way to measure how spread out a data set is. It provides information about how far away each data point is from the mean and sets the foundation for the standardisation of scores across a variety of topics.

Calculating the sample mean

When data is collected, it may be taken from a **population** (the entire group from which a conclusion can be drawn) or a **sample** (a smaller subset of the population). As data collected from a population is not always easily accessible, the scope of this course focuses on data collected from a sample. As such, when calculating the mean of a data set in this course, this is actually a reference to the mean of a sample (or sample mean), which is represented as $\bar{x}$.

The **mean** is a measure of centre that averages out all values into even groupings. It is calculated by adding all data values in a sample and then dividing the sum by the number of values. This can be expressed in the formula:

$\bar{x} = \frac{\Sigma x}{n}$, where Σx is the 'sum of all values', and n is the number of values in the data set.

Worked example 1

The amount of *money spent*, in dollars, of 10 customers at a supermarket is shown. Calculate the mean amount of *money spent.*

11.60 55.50 7.95 42.15 17.10 2.00 82.55 26.85 5.40 21.90

Explanation

Step 1: Calculate Σx and determine n.

$$\begin{aligned}\Sigma x &= \text{sum of all values}\\ &= 11.60 + 55.5 + 7.95 + 42.15 + 17.10 + 2.00\\ &\quad + 82.55 + 26.85 + 5.40 + 21.90\\ &= 273\end{aligned}$$

$$\begin{aligned}n &= \text{number of data values}\\ &= 10\end{aligned}$$

Step 2: Calculate the mean.

$$\begin{aligned}\bar{x} &= \frac{\Sigma x}{n}\\ &= \frac{273}{10}\\ \bar{x} &= 27.3\end{aligned}$$

Continues →

Answer

$27.30

Calculating the sample mean and standard deviation using technology

Once the mean has been calculated, the standard deviation can be used to give an indication of the spread of a data set. The standard deviation referenced in this course is for a sample and is represented as s_x. The **standard deviation** is a measure of spread that is based on the average deviation (or difference) of each data point compared to the mean. The standard deviation of a sample can be calculated manually with the formula

$$s_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}},$$

where $\Sigma(x - \bar{x})^2$ is the sum of the squared differences between each data point and the mean, and n is the number of values in the data set; however, it is more commonly calculated using technology rather than by hand.

Worked example 2

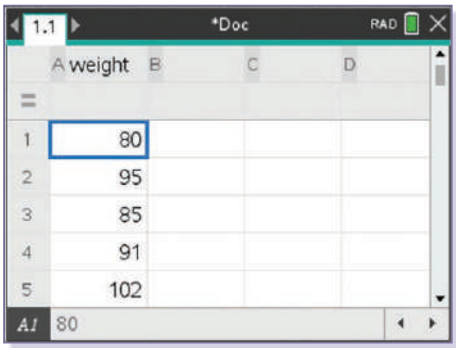
The *weight*, in kilograms, of a sample of 10 rugby players are shown. Determine the mean and standard deviation of the rugby players' *weight*, rounded to two decimal places.

80 95 85 91 102 93 87 78 84 90

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Name column A 'weight' and enter the data values starting from row 1 into the column below.

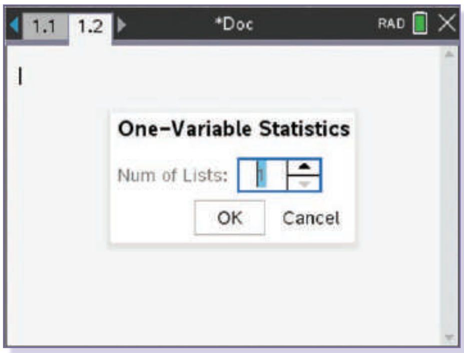


	A weight	B	C	D
=				
1	80			
2	95			
3	85			
4	91			
5	102			

A1 80

Step 3: Press ctrl + doc▾, and select '1: Add Calculator'.

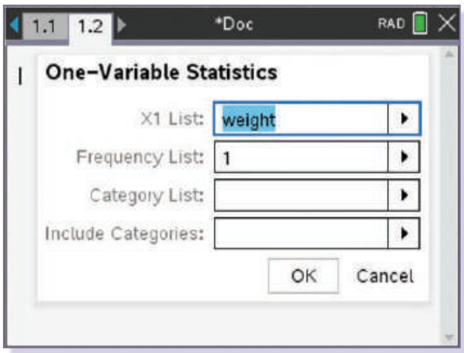
Step 4: Press menu. Select '6: Statistics' → '1: Stat Calculations' → '1: One-Variable Statistics'. As there is only one data set, on the screen that follows, select 'OK'.



One-Variable Statistics

Num of Lists: 1

Step 5: On the next screen, select 'weight' as the 'X1 List' using the dropdown list, then select 'OK'.



One-Variable Statistics

X1 List: weight

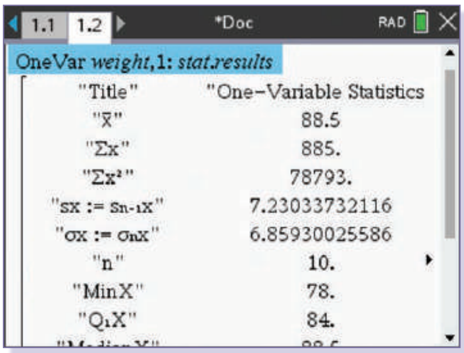
Frequency List: 1

Category List:

Include Categories:

Continues →

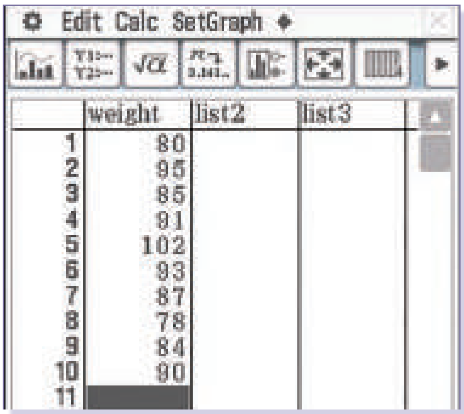
Step 6: Identify the sample mean, $\bar{x}$, and standard deviation, s_x, (scrolling up may be required).



```
OneVar weight,1: stat.results
   "Title"        "One-Variable Statistics
    "x̄"                 88.5
    "Σx"                885.
    "Σx²"              78793.
 "sx := sn-1x"     7.23033732116
 "σx := σnx"       6.85930025586
    "n"                  10.
   "MinX"                78.
   "Q₁X"                 84.
```

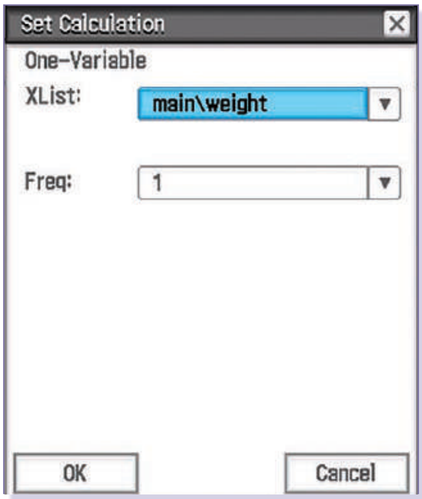
Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics. Name list1 'weight' and enter the data values starting from row 1 into the column below.



	weight	list2	list3
1	80		
2	95		
3	85		
4	91		
5	102		
6	93		
7	87		
8	78		
9	84		
10	90		
11			

Step 2: Tap the 'Calc' menu at the top of the screen and select 'One-Variable'. On the screen that follows, select 'main\weight' as the 'XList' using the dropdown list, then tap 'OK'.

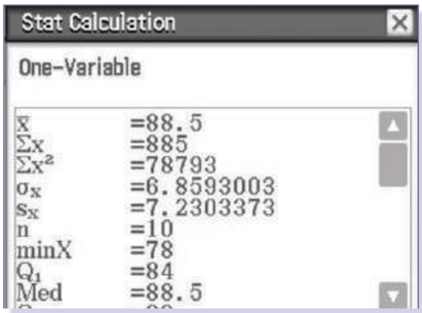


One-Variable

XList: main\weight

Freq: 1

Step 3: Identify the sample mean, $\bar{x}$,and standard deviation, s_x.



```
One-Variable

x̄       =88.5
Σx      =885
Σx²     =78793
σx      =6.8593003
sx      =7.2303373
n       =10
minX    =78
Q₁      =84
Med     =88.5
```

Answer – Method 1 and 2

Mean: 88.50 kg

Standard deviation: 7.23 kg

Exam question breakdown

VCAA 2017 Exam 1 Data analysis Q3

The following table shows the forearm *circumference*, in centimetres, of a sample of 10 people selected from a group of 252 people.

circumference	26.0	27.8	28.4	25.9	28.3	31.5	28.2	25.9	27.9	27.8

The mean, $\bar{x}$, and the standard deviation, s_x, of the forearm *circumference* for this sample of people are closest to

A. $\bar{x} = 1.58 \quad s_x = 27.8$

B. $\bar{x} = 1.66 \quad s_x = 27.8$

C. $\bar{x} = 27.8 \quad s_x = 1.58$

D. $\bar{x} = 27.8 \quad s_x = 1.66$

E. $\bar{x} = 27.8 \quad s_x = 2.30$

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Name column A 'circumf' and enter the data values starting from row 1 into the column below.

Step 3: Press ctrl + doc ▾, and select '1: Add Calculator'.

Step 4: Press menu. Select '6: Statistics' → '1: Stat Calculations' → '1: One-Variable Statistics'. As there is only one data set, on the screen that follows, select 'OK'.

Step 5: On the next screen, select 'circumf' as the 'X1 List' using the dropdown list, then select 'OK'.

Step 6: Identify the sample mean, $\bar{x}$, and standard deviation, s_x.

$\bar{x} = 27.8$ and $s_x = 1.66$

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics. Name list1 'circumf' and enter the data values starting from row 1 into the column below.

Step 2: Tap the 'Calc' menu at the top of the screen and select 'One-Variable'. On the screen that follows, select 'main\circumf' as the 'XList' using the dropdown list, then tap 'OK'.

Step 3: Identify the sample mean, $\bar{x}$, and standard deviation, s_x.

$\bar{x} = 27.8$ and $s_x = 1.66$

Answer - Method 1 and 2

D

88% of students answered this question correctly.

The most common incorrect responses were C, where students read the population standard deviation, σ_x, instead of the sample standard deviation, s_x, and B, where students read the mean and standard deviation in the wrong order.

1G Questions

Calculating the sample mean

1. A data set contains the following values.

35 21 59 43 56 30

The mean of the data set is closest to

A. 40.0 **B.** 40.6 **C.** 40.7 **D.** 41.0

2. Calculate the mean of the following data set.

141 159 130 164 150 147

3. Calculate the mean *number of pets* from the following data set, rounded to one decimal place.

number of pets	**frequency**
0	5
1	8
2	4
3	2
total	19

Calculating the sample mean and standard deviation using technology

4. For the following data set,

91 56 67 34 65 89 76 99 33 21 23 54 34 32 76 67 84

the sample mean and standard deviation are closest to

A. *mean* = 58.9 *standard deviation* = 24.6

B. *mean* = 58.9 *standard deviation* = 25.4

C. *mean* = 25.4 *standard deviation* = 58.9

D. *mean* = 58.8 *standard deviation* = 24.6

5. Calculate the sample mean and standard deviation for the following data set (to the nearest one decimal place).

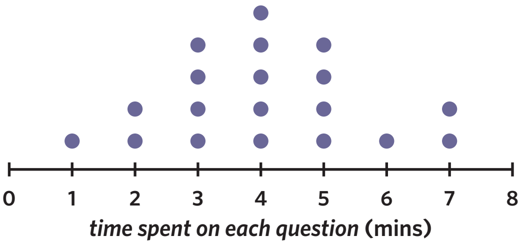
151.4 147.6 134.1 156.5 184.3 165.7 167.0 164.1 155.9 157.2

6. Calculate the sample mean and standard deviation for the following data set, rounded to one decimal place.

number of pens in a pencil case	0	1	2	3	4	5
frequency	3	10	5	4	1	1

Joining it all together

7. While studying for her geography exam, Layla recorded the *time spent on each question* in order to determine her speed. Her results are displayed in the dot plot, correct to the nearest minute.



a. Identify Σx and n for the data set.

b. Hence, calculate the mean amount of *time spent on each question*, rounded to the nearest minute.

c. Determine the standard deviation, rounded to two decimal places.

8. The following stem plot displays the *daily sales* at a crepe shop each day for two weeks.

Key: 3 | 1 = 31 sales

3	1
4	3
5	4 7
6	1 2 4
7	2 3 5 9
8	1 1 3

a. Identify Σx and n for the data set.

b. Hence, calculate the mean amount of *daily sales* at the crepe shop over the two weeks, rounded to the nearest whole number of sales.

c. Determine the standard deviation, rounded to two decimal places.

9. Ahava collects data from her colleagues on the number of *disposable coffee cups* used in a week. She is staggered to find that in total, her colleagues throw away 474 *disposable coffee cups* every week, with an average of 6 per person.

a. How many colleagues did Ahava ask?

b. Ahava wants to reduce the average number of disposable coffee cups that her colleagues use over the week. Although she would ideally like to eliminate their use altogether, she decides to start small and would like to lower the average coffee cups used by each person to 5 over the next week.

How many less coffee cups will need to be disposed of in the coming week to achieve this target?

Exam practice

10. In the sport of heptathlon, athletes compete in seven events.

These events are the 100 m hurdles, high jump, shot-put, javelin, 200 m run, 800 m run and long jump.

Fifteen female athletes competed to qualify for the heptathlon at the Olympic Games.

Their results for three of the heptathlon events – high jump, shot-put and javelin – are shown in Table 1.

Table 1

athlete number	1	2	3	4	5	6	7	8
high jump **(metres)**	1.76	1.79	1.83	1.82	1.87	1.73	1.68	1.82
shot-put **(metres)**	15.34	16.96	13.87	14.23	13.78	14.50	15.08	13.13
javelin **(metres)**	41.22	42.41	46.53	40.53	40.62	45.62	42.33	40.88

athlete number	9	10	11	12	13	14	15
high jump **(metres)**	1.83	1.87	1.87	1.80	1.83	1.87	1.78
shot-put **(metres)**	14.22	13.62	12.01	12.88	12.68	12.45	11.31
javelin **(metres)**	39.22	42.51	42.75	38.12	42.65	41.32	42.88

Complete Table 2 by calculating the mean height jumped for the *high jump*, in metres, by the 15 athletes. Write the answer in the space provided in the table. (1 MARK)

Table 2

statistic	***high jump*** **(metres)**	***shot-put*** **(metres)**
mean		13.74
standard deviation	0.06	1.43

VCAA 2021 Exam 2 Data analysis Q1b

87% of students answered this question correctly.

11. The *body density*, in kilograms per litre, and *weight*, in kilograms, of a sample of 12 orangutans are shown in the table.

***body density* (kg/litre)**	1.07	1.07	1.08	1.08	1.03	1.05	1.07	1.06	1.07	1.09	1.02	1.09
***weight* (kg)**	70.1	90.4	73.2	85.0	84.3	95.6	71.7	95.0	80.2	87.4	94.9	65.3

For these 12 orangutans, determine the mean of their *body density*, in kilograms per litre. (1 MARK)

Adapted from VCAA 2020 Exam 2 Data analysis Q4aii

77% of students answered this type of question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

12. The parallel boxplots shown display the data for twenty Year 12 students competing against twenty Year 10 students to see who can blow the most bubbles.

The five-number summary of the *number of bubbles* for Year 12 students is equal to:

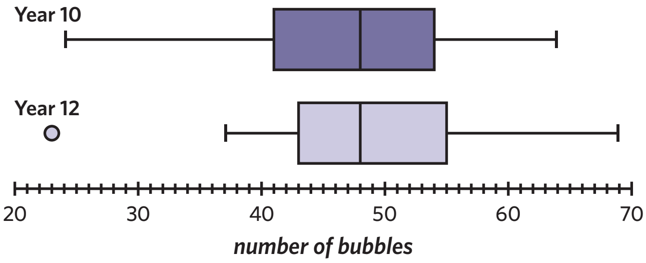
A. 23, 43, 48.5, 55, 69

B. 26, 41, 48, 54, 64

C. 23, 41, 48.5, 54, 69

D. 37, 43, 48, 54, 64

E. 37, 43, 48.5, 55, 69



Adapted from VCAA 2017 Exam 1 Data analysis Q2

Recursion and financial modelling *Year 11 content*

13. Wally decides that he needs to practise playing the piano. On the first day, he spends a total of two hours practising.

Each following day, he spends five minutes less than the previous day.

Let w_n be the number of minutes that Wally spends practising the piano on day n.

A recurrence relation that can be used to model this behaviour for $1 \leq n \leq 24$ is

A. $w_{n+1} = w_n + 5, \quad w_1 = 2$

B. $w_{n+1} = w_n \times 1.05, \quad w_1 = 120$

C. $w_{n+1} = w_n \times 0.95, \quad w_1 = 120$

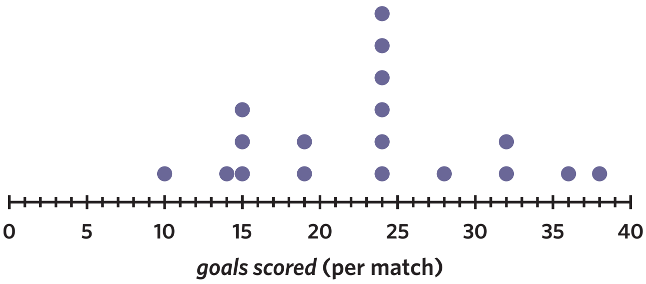
D. $w_{n+1} = w_n - 5, \quad w_1 = 2$

E. $w_{n+1} = w_n - 5, \quad w_1 = 120$

Adapted from VCAA 2014 Exam 1 Number patterns Q4

Data analysis *Year 11 content*

14. The following dot plot shows the distribution of *goals scored*, per match, throughout 18 matches in a local football league.



a. Write down the

i. range (1 MARK)

ii. median (1 MARK)

b. What is the number of goals scored at Q_3? (1 MARK)

Adapted from VCAA 2016 Exam 1 Data analysis Q1

1H The normal distribution

STUDY DESIGN DOT POINT

- the normal model for bell-shaped distributions and the use of the 68–95–99.7% rule to estimate percentages and to give meaning to the standard deviation; standardised values (z-scores) and their use in comparing data values across distributions

1A 1B 1C 1D 1E 1F 1G 1H 1I

KEY SKILLS

During this lesson, you will be:

- calculating proportions in normal distributions
- calculating values in normal distributions
- determining the mean and standard deviation of a normal distribution.

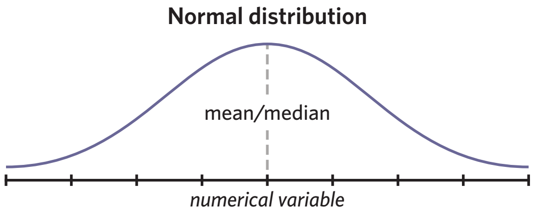
KEY TERMS

- Normal distribution

Data sets can take on a variety of different distributions and shapes. For data sets that exhibit a normal distribution, there are assumptions and generalisations that can be made that assist in making predictions about the data. These predictions make it possible to further analyse a data set by utilising unique properties of the mean and standard deviation of normal distributions.

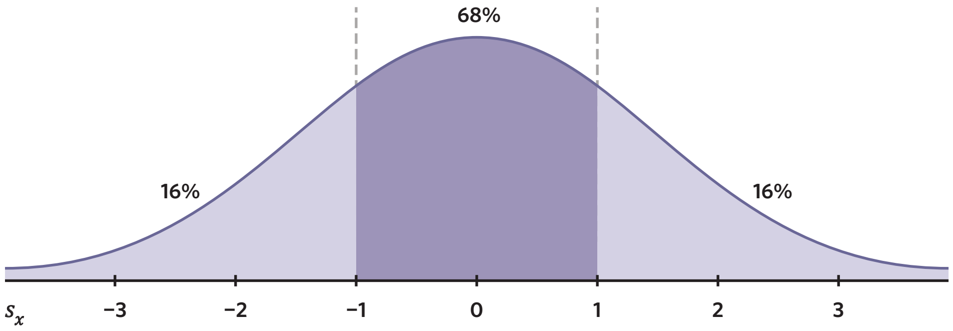
Calculating proportions in normal distributions

A **normal distribution** is a symmetrical (or approximately symmetrical) numerical data set that is centred around the mean, with a width determined by the standard deviation. Normal distributions are commonly known to be 'bell-shaped'. In a normal distribution, the mean and median are equal and are located through the central mirror line of the distribution. Some instances where data may be approximately normally distributed include study scores and IQ.

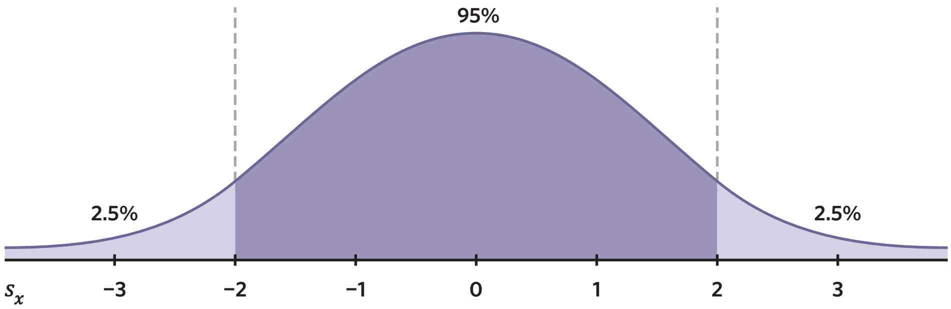


For a data set that is normally distributed, the data is spread around the mean with respect to the standard deviation. The 68–95–99.7% rule states that:

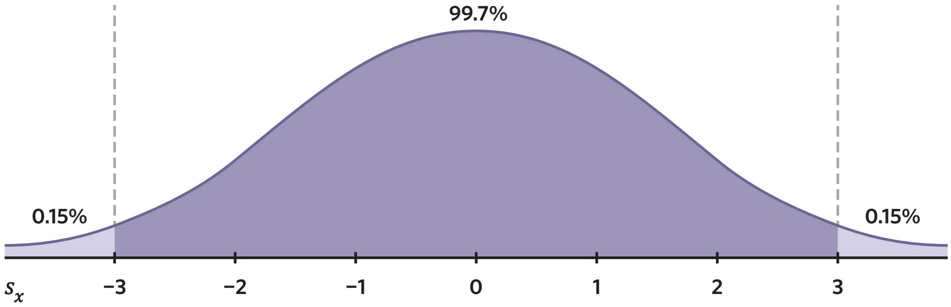
- 68% of data lies within one standard deviation (s_x) on either side of the mean ($\pm 1\ s_x$)



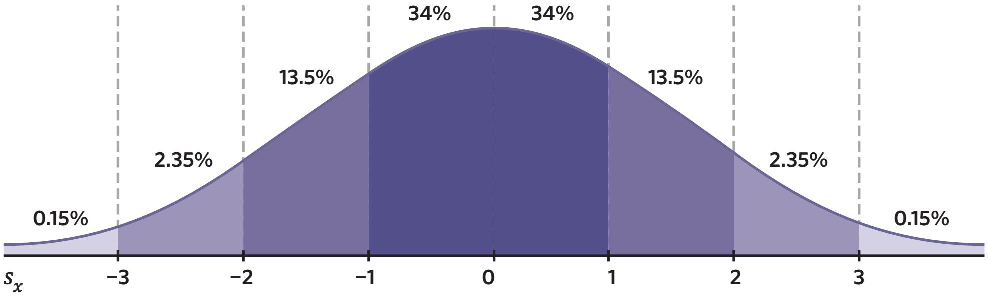
- 95% of data lies within two standard deviations on either side of the mean ($\pm 2\ s_x$)



- 99.7% of data lies within three standard deviations on either side of the mean ($\pm 3\ s_x$)



This allows a percentage breakdown of the data into groups based on the location of a data point with respect to the mean.



This concept can then be applied to calculate the proportion of data that lies between specified boundaries, as well as calculating the expected number within a sample that fulfil a specified criteria.

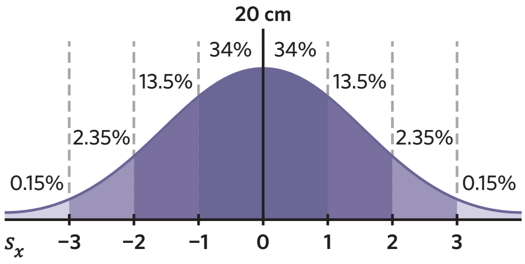
Worked example 1

An entire girl's basketball club decided to shave their heads to raise money for cancer. They also donated their hair to be made into wigs for those undergoing treatment. The *length of hair donated* was approximately normally distributed with a mean of 20 cm and a standard deviation of 4 cm.

a. What percentage of hair lengths donated were between 16 cm and 32 cm?

Explanation

Step 1: Locate and label the mean on the normal distribution graph.



Continues →

Step 2: Calculate and label the values one, two and three standard deviations on either side of the mean.

1 standard deviation:

$20 - 1 \times 4 = 16$

$20 + 1 \times 4 = 24$

2 standard deviations:

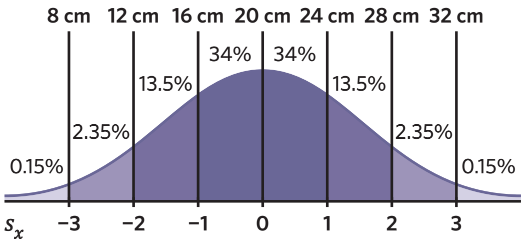
$20 - 2 \times 4 = 12$

$20 + 2 \times 4 = 28$

3 standard deviations:

$20 - 3 \times 4 = 8$

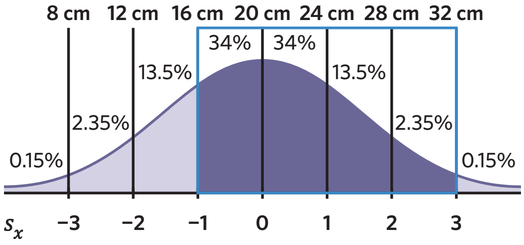
$20 + 3 \times 4 = 32$



Step 3: Identify the boundaries stated in the question.

The lower boundary is 16 cm.

The upper boundary is 32 cm.



Step 4: Calculate the sum of the percentages within the boundaries.

$34 + 34 + 13.5 + 2.35 = 83.85$

Answer

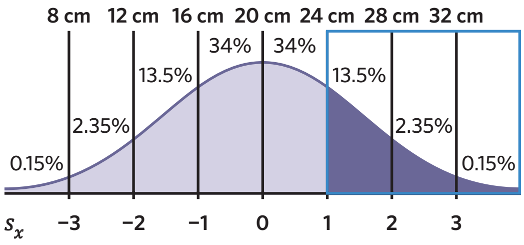
83.85%

b. Donated hair lengths greater than 24 cm are highly sought after as they can be customised for a diverse range of hairstyles. What percentage of the girl's basketball club donated hair greater than 24 cm?

Explanation

Step 1: Determine the boundary, and the equivalent number of standard deviations.

A lower boundary of 24 cm is equivalent to 1 standard deviation greater than the mean.



Step 2: Calculate the percentage of data that fulfils the criteria.

$13.5 + 2.35 + 0.15 = 16$

Answer

16%

c. If there are 30 girls at the basketball club, how many of them are expected to have donated hair lengths longer than 24 cm, rounded to the nearest person?

Explanation

From part **b**, 16% of the hair lengths donated were longer than 24 cm.

Therefore, calculate 16% of 30 girls.

$0.16 \times 30 = 4.8$

Answer

5 girls

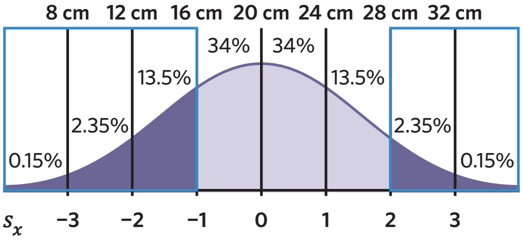
Continues →

d. How many girls are expected to have donated hair lengths that were either less than 16 cm or greater than 28 cm, rounded to the nearest person?

Explanation

Step 1: Using the labelled normal distribution graph, identify the boundaries stated in the question.

There are two separate boundary conditions that do not overlap.



Step 2: Calculate the sum of the percentages within the boundaries.

$0.15 + 2.35 + 13.5 + 2.35 + 0.15 = 18.5\%$

Step 3: Calculate the expected value.

18.5% of 30 girls.

$0.185 \times 30 = 5.55$

Answer

6 girls

Calculating values in normal distributions

Data sets that exhibit a normal distribution can also be used to determine the values within which a certain proportion of the data will lie.

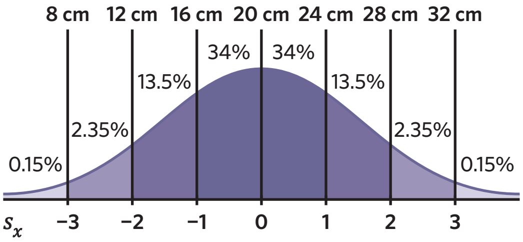
Worked example 2

An entire girl's basketball club decided to shave their heads to raise money for cancer. They also donated their hair to be made into wigs for those undergoing treatment. The *length of hair donated* was approximately normally distributed with a mean of 20 cm and a standard deviation of 4 cm.

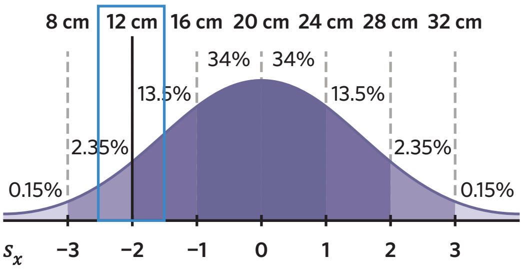
a. 97.35% of the hair lengths will fall within 12 cm and which other value?

Explanation

Step 1: Identify and label the mean, and the values that lie 1, 2, and 3 standard deviations on either side of the mean.



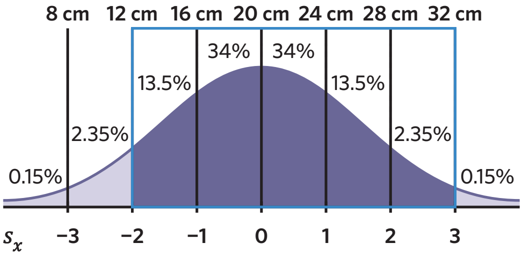
Step 2: Locate the boundary stated in the question.



Step 3: Determine the second boundary.

As 97.35% of the data does not lie below 12 cm, add percentages that are greater than 12 cm until 97.35% is obtained.

$13.5 + 34 + 34 + 13.5 + 2.35 = 97.35$



Continues →

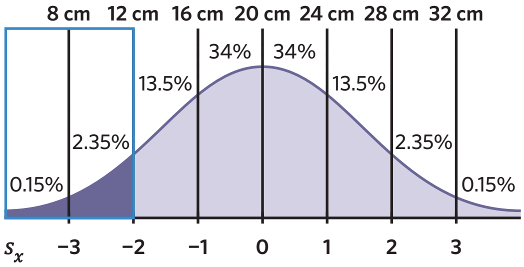
Answer

32 cm

b. 2.5% of hair donated by the team will be disposed of, as it falls below a required minimum length. What is this required minimum length?

Explanation

Using the labelled normal distribution graph, identify the appropriate boundary.



0.15 + 2.35 = 2.5%

Therefore, 2.5% of hair lengths are below 12 cm.

Answer

12 cm

Determining the mean and standard deviation of a normal distribution

The 68–95–99.7% rule can also be used to calculate the mean and standard deviation of a data set if unknown.

Worked example 3

Duncan's light bulb factory produces light bulbs.

The *lifetime* of their light bulbs are approximately normally distributed. It is known that approximately 95% of the light bulbs have a *lifetime* ranging from 800 hours to 1600 hours.

a. Calculate the standard deviation for the *lifetime* of a light bulb.

Explanation

Step 1: Determine the number of standard deviations.

95% of data lies between two standard deviations either side of the mean. This is a total of four standard deviations.

Step 2: Calculate the value of one standard deviation.

Four standard deviations have a value of 1600 − 800 = 800.

Therefore, one standard deviation is equal to 800 ÷ 4 = 200.

Answer

200 hours

Continues →

b. Calculate the mean *lifetime* of a light bulb.

Explanation

The mean will lie two standard deviations greater than the lower boundary or two standard deviations less than the upper boundary.

$$\begin{aligned} 800 + 2 \times 200 &= 800 + 400 \\ &= 1200 \end{aligned}$$

Answer

1200 hours

Exam question breakdown

Adapted from VCAA 2021 Exam 2 Data analysis Q1d

In a qualifying competition, the heights jumped in the high jump are expected to be approximately normally distributed.

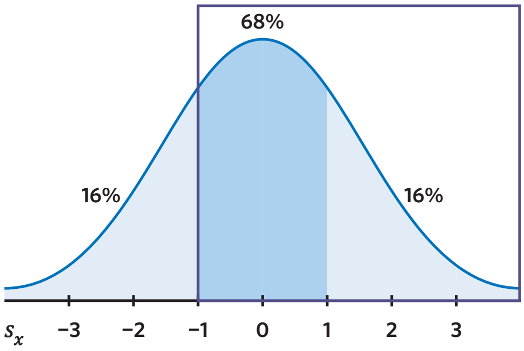
Chara's jump in this competition places her one standard deviation below the mean.

Use the 68–95–99.7% rule to calculate the percentage of athletes who would be expected to jump higher than Chara in the qualifying competition. (1 MARK)

Explanation

Step 1: Identify the boundaries using the normal distribution graph.

The question places Chara's jump one standard deviation below the mean (–1) and asks for athletes that have jumped higher.



Answer

84%

Step 2: Calculate the sum of the percentages within the boundaries.

$68 + 16 = 84$

60% of students answered this type of question correctly.

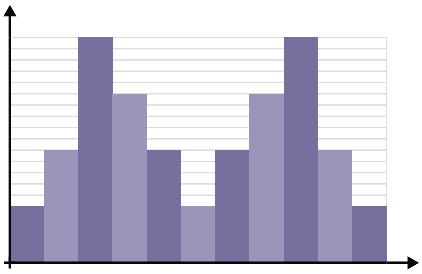
The most common incorrect answer was 16%. This is most likely caused by students finding the percentage of athletes who were expected to jump lower than Chara, rather than higher.

1H Questions

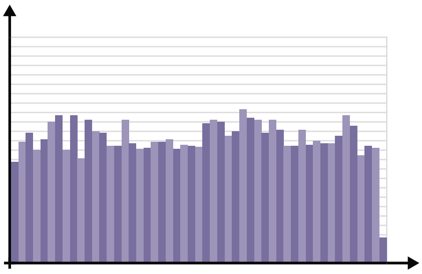
Calculating proportions in normal distributions

1. Which of the following is an example of a normal distribution?

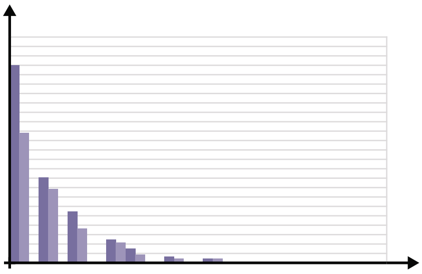
A.



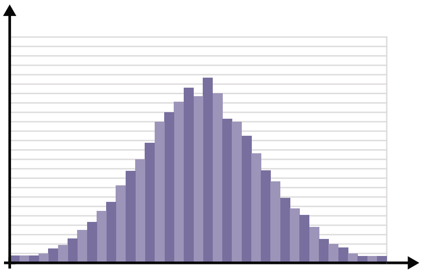
B.



C.



D.



2. 3000 students were surveyed about the *volume of milk* (mL) they add to their breakfast tea each morning. The results were approximately normally distributed with a mean of 30 mL and a standard deviation of 5 mL. What percentage of students add a *volume of milk* that is:

a. less than 25 mL?

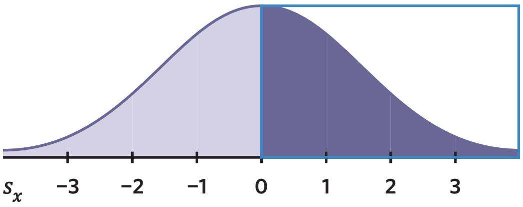
b. either less than 20 mL or greater than 40 mL?

3. Malia goes to a turkey farm before Christmas to have a look at their range of turkeys. She finds that the distribution of *weight* for the male turkeys is approximately normally distributed with a mean of 8 kilograms and a standard deviation of 1 kilogram. Malia has all of her extended family coming for Christmas lunch this year and will need a turkey that weighs more than 11 kilograms. If there are 40 000 turkeys available on the day that she visits the farm, how many turkeys will be suitable for her?

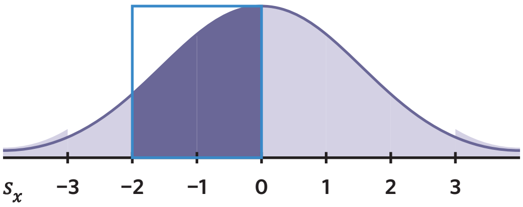
Calculating values in normal distributions

4. Which of the following graphs identifies 49.85% of data?

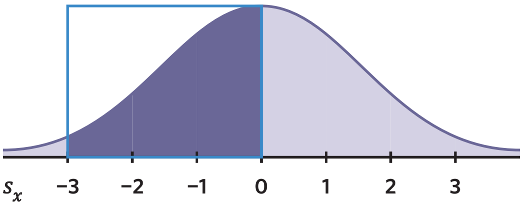
A.



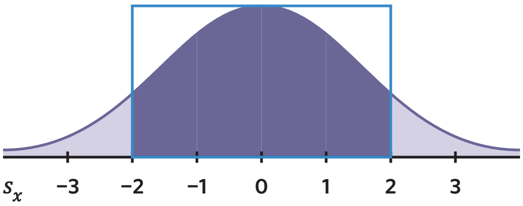
B.



C.



D.



5. The *weight*, in kilograms, of mountain zebras is approximately normally distributed with a mean of 280 kg and a standard deviation of 21 kg.

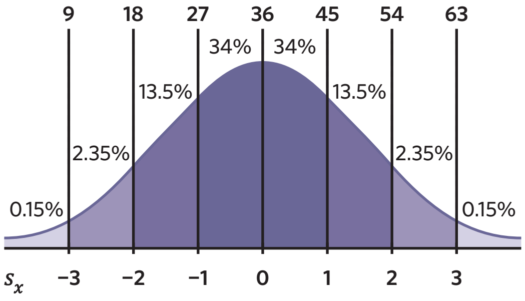
   a. Approximately 16% of mountain zebras will weigh more than what weight?

   b. Approximately 81.5% of mountain zebras will weigh between 259 kg and what other weight?

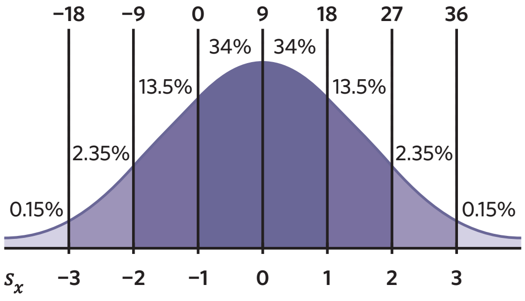
Determining the mean and standard deviation of a normal distribution

6. Which of the following graphs shows a normal distribution with a mean of 63 and a standard deviation of 9?

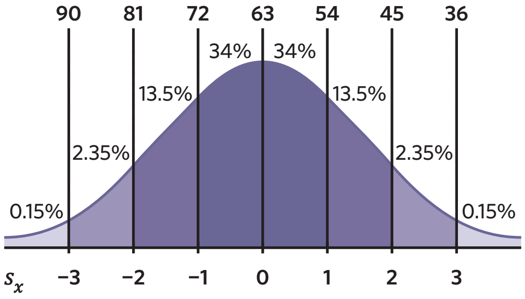
A.



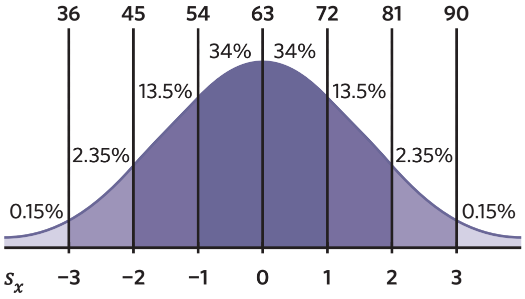
B.



C.



D.



7. A manufacturing company produces plastic bottles from recycled plastics that have been collected from waterways. The *density*, in g/cm^3, of each bottle is approximately normally distributed. The company claims that their manufacturing process is so consistent that 99.7% of the bottles produced have a *density* ranging between 0.92 and 1.16 g/cm^3.

   a. Calculate the standard deviation for the *density* of the plastic bottles produced at the manufacturing company.

   b. Calculate the mean *density* of the plastic bottles.

Joining it all together

8. The heights of 2000 seedlings planted by Michelle in her garden last year were measured and found to be approximately normally distributed. The mean of the heights was found to be 140 cm and the standard deviation was 12 cm.

   a. What percentage of Michelle's plants were between

      i. 128 cm and 140 cm?

      ii. 104 cm and 116 cm?

   b. Michelle is a diligent gardener and knows that she needs to harvest any plants that grow taller than 176 cm. Approximately how many plants will she need to harvest from her garden?

   c. The smallest 2.5% of seedlings will need extra care if they are to grow to their full potential. What height would the tallest of these seedlings be?

9. It is known that travel time from home to school for students at Schoolrolo Secondary College is approximately normally distributed with a mean of 45 minutes and a standard deviation of 10 minutes.

   a. Find the percentage of students who take less than 55 minutes to arrive at school.

b. Given that Schoolrolo Secondary College has 1800 students, find the approximate number of students who take more than 75 minutes to arrive at school. Round to the nearest number of students.

c. Find a time interval in which approximately 95% of the students will take to arrive at school.

10. The *mean time before failure* (*MTBF*) of an aeroplane engine is approximately normally distributed with a mean of 25 000 hours. Assume that one airline has a fleet of 200 aircrafts with each plane having two engines.

a. How many engines does the airline have in total?

Approximately 380 of these engines will have an *MTBF* between 24 000 and 26 000 hours.

b. Calculate the standard deviation.

c. Find the number of engines that have an *MTBF* more than 25 500 hours.

d. State a time interval in which approximately 68% of the engines will fail.

Exam practice

11. The time taken to *travel* between two regional cities is approximately normally distributed with a mean of 70 minutes and a standard deviation of 2 minutes.

The percentage of *travel* times that are between 66 minutes and 72 minutes is closest to

A. 2.5%
B. 34%
C. 68%
D. 81.5%
E. 95%

VCAA 2019 Exam 1 Data analysis Q6

79% of students answered this question correctly.

12. The *temperature difference* between the *minimum daily temperature* and the *maximum daily temperature* in November 2017 at a location is approximately normally distributed with a mean of 9.4 °C and a standard deviation of 3.2 °C.

Determine the number of days in November 2017 for which this *temperature difference* is expected to be greater than 9.4 °C. (1 MARK)

VCAA 2019 Exam 2 Data analysis Q2b

58% of students answered this question correctly.

13. In a large population of moths, the number of eggs per cluster is approximately normally distributed with a mean of 165 eggs and a standard deviation of 25 eggs.

Using the 68–95–99.7% rule, determine

a. the percentage of clusters expected to contain more than 140 eggs (1 MARK)

b. the number of clusters expected to have less than 215 eggs in a sample of 1000 clusters. (1 MARK)

VCAA 2017 Exam 2 Data analysis Q1b

Part **a**: **69%** of students answered this question correctly.
Part **b**: **47%** of students answered this question correctly.

Questions from multiple lessons

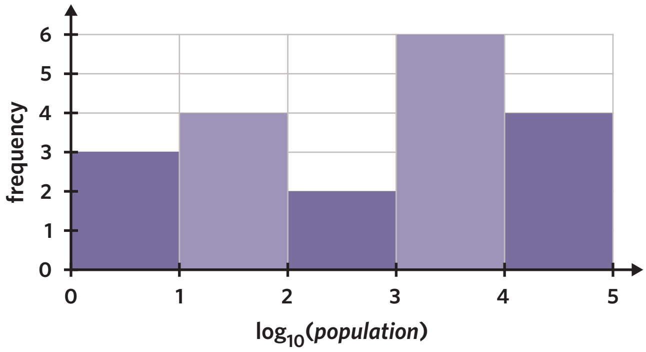
Data analysis

14. The following histogram displays the distribution of the $\log_{10}(population)$ of 19 planets in a faraway galaxy.

The median population is between

A. 10 and 100
B. 100 and 1000
C. 1000 and 10 000
D. 2 and 3
E. 3 and 4

Adapted from VCAA 2017 Exam 1 Data analysis Q4



Recursion and financial modelling *Year 11 content*

15. A dress shop is having a closing-down sale and is selling the 1800 dresses it has in stock.

On the first day of the sale, 100 dresses are sold.

On the second day, 125 dresses are sold.

On the third day, 150 dresses are sold.

This pattern continues until all 1800 dresses have been sold.

How long does it take to sell all of the dresses?

A. 8 days **B.** 9 days **C.** 18 days **D.** 19 days **E.** 67 days

Adapted from VCAA 2013 Exam 1 Number patterns Q6

Data analysis *Year 11 content*

16. A study was conducted investigating the relationship between *sleep*, measured in hours, and *blood pressure*, measured in mmHg. The following least squares regression equation was obtained.

$\textit{blood pressure} = 220 - 10.47 \times \textit{sleep}$

a. Which variable is the response variable? (1 MARK)

b. Interpret the slope of the regression equation in terms of the given variables. (2 MARKS)

Adapted from VCAA 2017NH Exam 2 Data analysis Q2bi,ii

1I *z*-scores

STUDY DESIGN DOT POINT

- the normal model for bell-shaped distributions and the use of the 68–95–99.7% rule to estimate percentages and to give meaning to the standard deviation; standardised values (*z*-scores) and their use in comparing data values across distributions

1A 1B 1C 1D 1E 1F 1G 1H **1I**

KEY SKILLS

During this lesson, you will be:

- calculating standardised (z) scores
- calculating actual scores from standardised (z) scores
- using standardised (z) scores to interpret data.

KEY TERMS

- Standardised score (z-score)
- Actual score

Comparing several data sets that exhibit a normal distribution can be difficult when the number of data values in the samples differ, especially if by a large amount. Standardised scores (or z-scores) are an important tool that statisticians use to compare normal distributions. Standardised scores extend on the 68–95–99.7% rule and provide a more precise measure of the location of each data value within a sample.

Calculating standardised (z) scores

A **standardised score**, also known as a **z-score**, is a measure of the number of standard deviations between the mean and a data value. When performing standardised score calculations, each data value in a data set is referred to as an ‘**actual score**’.

Standardised scores can be:

- positive, indicating that the actual score is above the mean
- zero, indicating that the actual score is equal to the mean
- negative, indicating that the actual score is below the mean.

A standardised score is calculated by subtracting the mean from the actual score, and then dividing the result by the standard deviation.

$$z = \frac{x - \bar{x}}{s_x}$$

- z is the standardised score
- x is the actual score
- $\bar{x}$ is the mean
- s_x is the standard deviation

Worked example 1

The weights of puppies within a litter are approximately normally distributed with a mean of 4 kg and a standard deviation of 0.3 kg.

Determine the z-score of a puppy that weighs 4.6 kg.

Continues →

Explanation

Step 1: Identify the mean, standard deviation and actual score.

$\overline{x} = 4$

$s_x = 0.3$

$x = 4.6$

Step 2: Calculate the z-score.

Substitute the values into the formula $z = \frac{x - \overline{x}}{s_x}$.

$$z = \frac{4.6 - 4}{0.3} = \frac{0.6}{0.3} = 2$$

Answer

2

Calculating actual scores from standardised (z) scores

A standardised score can also be used to calculate the actual score using the formula:

$x = \overline{x} + (z \times s_x)$

Worked example 2

The number of black stripes on tigers at the Royal Melbourne Zoo is known to be normally distributed with a mean of 55 and a standard deviation of 2. One of the tigers has a standardised number of stripes of $z = -2.5$. How many stripes does the tiger actually have?

Explanation

Step 1: Identify the mean, standard deviation and standardised score.

$\overline{x} = 55$

$s_x = 2$

$z = -2.5$

Step 2: Calculate the actual score.

Substitute the values into the formula $x = \overline{x} + (z \times s_x)$.

$$x = 55 + (-2.5 \times 2) = 55 - 5 = 50$$

Answer

50 stripes

Using standardised (z) scores to interpret data

Comparing raw data values across multiple data sets may be misleading. Standardised scores help to look at data in a way which allows for comparison between different data sets.

For example, a test score of 70 might be considered low for an easy test in which everyone else did well on, but high for a more difficult test, which people struggled with.

The standardised score takes into account the mean and standard deviation of each data set and, when paired with the 68–95–99.7% rule, allows for a better comparison to be made.

Worked example 3

The results for a Chemistry exam were approximately normally distributed, with a mean of 71 and a standard deviation of 3. The results for a Physics exam were also approximately normally distributed, with a mean of 95 and a standard deviation of 6.

a. A student taking both Chemistry and Physics scored 77 in her final exam for both subjects. Did she do equally well in both exams in relation to her peers? Use standardised scores and the 68–95–99.7% rule to justify your answer.

Continues →

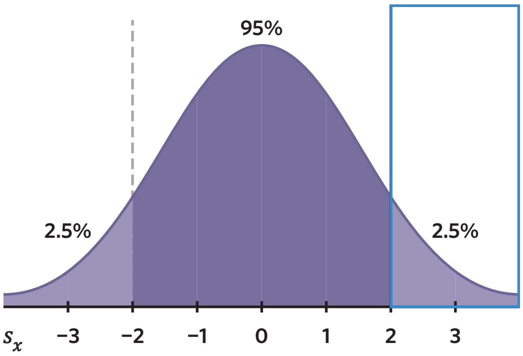
Explanation

Step 1: Calculate the standardised score for Chemistry.

$$z = \frac{77 - 71}{3}$$
$$= 2$$

Step 2: Interpret the standardised score for Chemistry using the 68–95–99.7% rule.

A *z*-score of 2 indicates that the student is two standard deviations above the mean Chemistry exam score. This means she is in the top 2.5% of the Chemistry class.

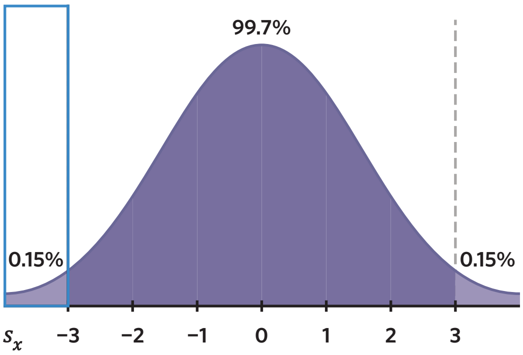


Step 3: Calculate the standardised score for Physics.

$$z = \frac{77 - 95}{6}$$
$$= -3$$

Step 4: Interpret the standardised score for Physics using the 68–95–99.7% rule.

A *z*-score of −3 indicates that the student is three standard deviations below the mean Physics exam score. This means she is in the bottom 0.15% of the Physics class.



Step 5: Write a worded answer.

Ensure the question is answered directly. Justify your answer using the information from the previous steps.

Answer

In Chemistry, the student's standardised score was $z = 2$ placing her in the top 2.5% of students.
In Physics, the student's standardised score was $z = -3$ placing her in the bottom 0.15% of students.
Therefore, she did not do equally well in both exams when compared to her peers.

b. Another student scores 66 on his Chemistry exam. Use standardised scores and the 68–95–99.7% rule to compare this student's score relative to the class.

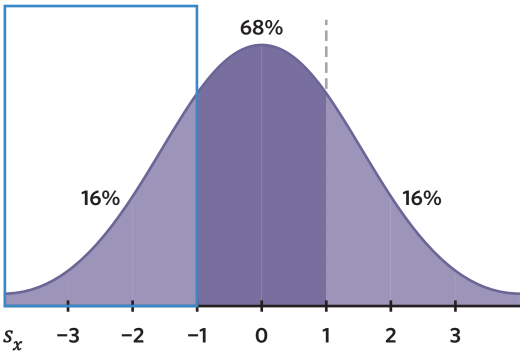
Explanation

Step 1: Calculate the standardised score for Chemistry.

$$z = \frac{66 - 71}{3}$$
$$= -1.666...$$

Step 2: Interpret the standardised score for Chemistry using the 68–95–99.7% rule.

A *z*-score of −1.67 indicates that the student is between one and two standard deviations below the mean Chemistry exam score. This means he is in the bottom 16% of the Chemistry class.



Continues →

Step 3: Write a worded answer.

Ensure the question is answered directly.
Justify your answer using the information from the previous steps.

Answer

In Chemistry, the student's standardised score was $z = -1.67$ placing him in the bottom 16% of students.

Exam question breakdown

VCAA 2016 Exam 1 Data analysis Q5

The weights of male players in a basketball competition are approximately normally distributed with a mean of 78.6 kg and a standard deviation of 9.3 kg.

Brett and Sanjeeva both play in the basketball competition.

When the weights of all players in the competition are considered, Brett has a standardised weight of $z = -0.96$ and Sanjeeva has a standardised weight of $z = -0.26$.

Which one of the following statements is **not** true?

A. Brett and Sanjeeva are both below the mean weight for players in the basketball competition.

B. Sanjeeva weighs more than Brett.

C. If Sanjeeva increases his weight by 2 kg, he would be above the mean weight for players in the basketball competition.

D. Brett weighs more than 68 kg.

E. More than 50% of the players in the basketball competition weigh more than Sanjeeva.

Explanation

To solve this question, check whether each option is true or false.

A: This is true. As the z-scores for both Brett and Sanjeeva are negative, this indicates that their weights are both below the mean weight. ✗

B: This is true. As Sanjeeva's z-score is higher than Brett's, this indicates that Sanjeeva weighs more than Brett. ✗

C: This is false.

Sanjeeva's actual weight:

$$x = \bar{x} + (z \times s_x)$$
$$= 78.6 + (-0.26 \times 9.3)$$
$$= 76.182$$

Add 2 kg to Sanjeeva's actual weight.

$76.182 + 2 = 78.182$

78.182 is less than the mean of 78.6. ✓

Answer

C

D: This is true.

Brett's actual weight:

$$x = \bar{x} + (z \times s_x)$$
$$= 78.6 + (-0.96 \times 9.3)$$
$$= 69.672$$ ✗

E: This is true. As Sanjeeva's z-score is negative, this indicates that her weight is below the mean weight, and therefore less than 50% of the competition's weight. ✗

58% of students answered this question correctly.

Students who did not answer this question correctly would have found the question difficult to navigate, and probably did not take a methodical approach. As the possible responses are quite word-heavy, students most likely were overwhelmed with the choices.

1I Questions

Calculating standardised (z) scores

1. A standardised score of $z = -2.4$ indicates an actual score is

A. less than 2 standard deviations greater than the mean.

B. more than 2 standard deviations greater than the mean.

C. equal to the mean.

D. more than 2 standard deviations less than the mean.

2. A school hockey team underwent a cardiovascular fitness test. The heart rates of the students (in beats per minute, bpm) were measured, and found to be approximately normally distributed, with a mean of 65 bpm and a standard deviation of 8 bpm.

Calculate the standardised heart rate, correct to two decimal places, for a student whose actual heart rate was

a. 81 bpm. **b.** 77 bpm. **c.** 63 bpm. **d.** 54 bpm.

3. A PE class played soccer every lesson for one term. Over this period, the number of goals kicked by each student was recorded and was found to be approximately normally distributed, with a mean of 13 and a standard deviation of 3.

Calculate a student's standardised score, correct to three significant figures, given that they kicked

a. 7 goals. **b.** 19 goals. **c.** 15 goals. **d.** 6 goals. **e.** 2 goals.

Calculating actual scores from standardised (z) scores

4. Which of the following actual scores will give a standardised score of zero in a data set with a mean of 52.0 and a standard deviation of 3.4?

A. 0.0 **B.** 3.4 **C.** 52.0 **D.** 62.2

5. The weights of a group of cats are approximately normally distributed with a mean weight of 3.5 kg and a standard deviation of 0.6 kg.

Correct to two decimal places, calculate the actual weight that corresponds to each of the following standardised weights.

a. 1.5 **b.** -0.8 **c.** -2.5

6. After a local survey, the annual wage of workers in a small town was found to be approximately normally distributed. The workers had a mean wage of $62 377 with a standard deviation of $7256. Calculate a worker's actual wage, correct to the nearest dollar, for the following standardised scores.

a. $z = 1.00$ **b.** $z = -1.50$ **c.** $z = 2.50$ **d.** $z = 1.74$ **e.** $z = -2.83$

Using standardised (z) scores to interpret data

7. Which of the following scenarios will most likely result in equal z-scores?

A. One student scoring 54% on two separate tests.

B. Two students scoring 54% on the same test in the same class.

C. Two students scoring 54% on the same test in different classes.

D. One student scoring 54% on one test and 45% on another test.

8. Matt's height is measured at 182.2 cm. In Australia, male heights are approximately normally distributed with a mean of 175.6 cm and a standard deviation of 6.6 cm. In the Netherlands, male heights are also approximately normally distributed with a mean of 183.8 cm and a standard deviation of 7.1 cm.

Use standardised scores and the 68–95–99.7% rule to compare Matt's height relative to males in both countries.

9. The men's 100 m sprint results are approximately normally distributed. Times continue to drop every year. The mean times and standard deviations, in seconds, in the years 2017, 2018 and 2019 are shown in the table.

year	**mean**	**standard deviation**
2017	10.43	0.09
2018	10.17	0.23
2019	9.91	0.10

Kamil has also seen a reduction in his times over the three years. In 2017, he recorded a best time of 10.55 seconds. In 2018, his personal best had dropped to 10.12 seconds. In 2019, it had dropped further to 9.97 seconds.

In which year did Kamil run fastest when compared against other athletes? Use standardised scores and the 68–95–99.7% rule to justify your answer.

Joining it all together

10. Suppose that the height of 18-year-old girls is normally distributed with a mean of 169.0 cm and a standard deviation of 2.6 cm.

a. Camilla is 18 years old and 164.0 cm tall. What is her standardised height, correct to two significant figures?

b. Jing is 18 years old and has a standardised height of $z = -0.65$. What is her actual height, correct to four significant figures?

c. Amara is 18 years old and has a standardised height of $z = 2$.

i. What is Amara's actual height?

ii. What percentage of 18 year old girls are shorter than Amara?

11. Two classes sat tests in both Chemistry and Physics. The tests were marked out of 120. Table 1 shows the mean and standard deviation of both classes for these tests. One value is missing.

Table 1

	class A		**class B**	
	$\overline{x}$	s_x	$\overline{x}$	s_x
Chemistry	79	5	83	
Physics	85	7	65	12

Isaac and Mohit are in class A, whilst Sarika is in class B. Table 2 shows their results.

Table 2

	Isaac	**Mohit**	**Sarika**
Chemistry	92	89	76
Physics	92	81	81

a. Use standardised scores to compare Isaac's performance, relative to his class, in both subjects.

b. Use standardised scores to compare Mohit and Sarika's performances in Physics, relative to their respective classes.

c. What percentage of class A did better than Mohit in Chemistry?

d. Sarika has a standardised score of $z = -1$ in Chemistry, relative to her class. What is the standard deviation of class B in Chemistry?

e. What percentage of class B did worse than Sarika in Chemistry?

f. The pass mark for the Chemistry test in class B was such that only 50% of the class passed the test. What percentage of the class did better than Sarika but still failed the test?

Exam practice

12. The pulse rates of a population of Year 12 students are approximately normally distributed with a mean of 69 beats per minute and a standard deviation of 4 beats per minute.

A student selected at random from this population has a standardised pulse rate of $z = -2.5$.

This student's actual pulse rate is

A. 59 beats per minute.

B. 63 beats per minute.

C. 65 beats per minute.

D. 73 beats per minute.

E. 79 beats per minute.

VCAA 2018 Exam 1 Data analysis Q3

86% of students answered this question correctly.

13. In the sport of heptathlon, athletes compete in seven events.

These events are the 100 m hurdles, high jump, shot-put, javelin, 200 m run, 800 m run and long jump. In shot-put, athletes throw a heavy spherical ball (a shot) as far as they can.

Fifteen female athletes competed to qualify for the heptathlon at the Olympic Games.

For shot-put, athlete number six, Jamilia, threw the shot 14.50 m.

The table shows the mean and standard deviation of the shot-put results for the 15 athletes.

statistic	***shot-put* (metres)**
mean	13.74
standard deviation	1.43

Calculate Jamilia's standardised score (z).

Round your answer to one decimal place. (1 MARK)

VCAA 2021 Exam 2 Data analysis Q1c

78% of students answered this question correctly.

14. In a large population of moths, the number of eggs in a cluster of moth eggs is approximately normally distributed with a mean of 165 eggs and a standard deviation of 25 eggs.

The standardised number of eggs in one cluster is given by $z = -2.4$.

Determine the actual number of eggs in this cluster. (1 MARK)

VCAA 2017 Exam 2 Data analysis Q1c

67% of students answered this question correctly.

15. The temperatures over a number of days in a city have a mean of 3.5 °C and a standard deviation of 3 °C.

The temperature for one of the data points is found to be −4.3 °C.

The standardised value of this temperature is

A. -4.3

B. -2.6

C. 2.6

D. 3.5

E. 4.3

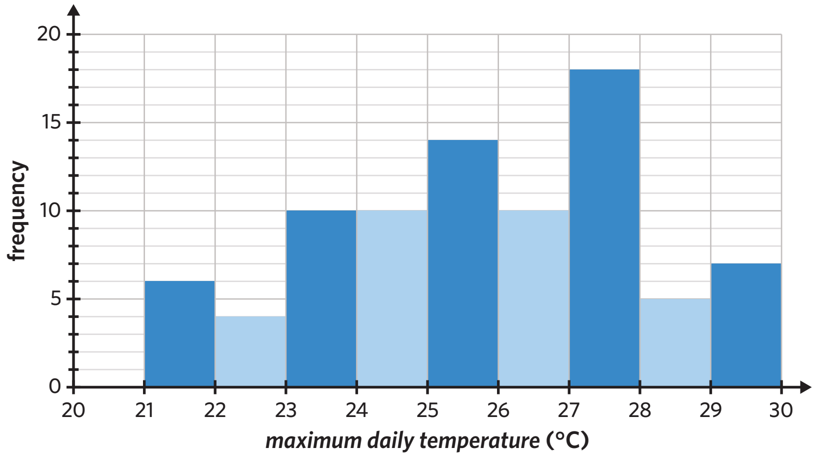
Adapted from VCAA 2017 Exam 1 Data analysis Q10

58% of students answered this type of question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

16. The following histogram displays the *maximum daily temperature* (°C) in a sample of 84 cities on a particular day.



The interquartile range for this distribution is closest to

A. 3 °C **B.** 4 °C **C.** 5 °C **D.** 7 °C **E.** 9 °C

Adapted from VCAA 2018NH Exam 1 Data analysis Q4

Recursion and financial modelling *Year 11 content*

17. Lorenzo bought an exotic pet python with a retail price of $6000.

However, he did not have enough money upfront. He paid a deposit of $2000 and repaid the rest of the balance with 36 monthly repayments of $200.

How much interest was Lorenzo charged?

A. $1200 **B.** $2200 **C.** $3200 **D.** $5200 **E.** $9200

Adapted from VCAA 2015 Exam 1 Business-related mathematics Q5

Data analysis *Year 11 content*

18. Students in a maths class recorded their heights and used the results to construct a five-number summary, as shown in the table. The distribution contains no outliers.

	minimum	Q_1	median	Q_3	maximum
***height* (cm)**	161	168	173	178	182

a. Construct a boxplot displaying the students' heights. (1 MARK)

b. What percentage of students had a height of 178 cm or less? (1 MARK)

Adapted from VCAA 2016 Exam 2 Data analysis Q2ai,ii