Chapter 1

Investigating data distributions

Chapter questions

- What is categorical data?
- What is numerical data?
- How do we summarise and display categorical data?
- How do we use the distribution of a categorical variable to answer statistical questions?
- What is a dot plot?
- What is a stem plot?
- What is a histogram?
- What is a boxplot?
- What are summary statistics, and how do we choose which ones to use?
- How do we use the distribution of a numerical variable to answer statistical questions?
- What is the normal distribution?
- What is the 68-95-99.7% rule and why is it useful?
- What are standardised values and why are they useful?

We can think of data as factual information about a person, object or situation which has been collected and recorded. In General Mathematics Units 1&2 we learned a range of statistical procedures to help us analyse such sets of data. In this chapter, we will review and extend our knowledge of those procedures for data which has been collected from a **single variable**, which is called **univariate** data.

ISBN 978-1-009-11041-9

Cambridge University Press

1A Types of data

Learning intentions

- To be able to classify data as categorical or numerical.
- To be able to further classify categorical data as nominal or ordinal.
- To be able to further classify numerical data as discrete or continuous.

A group university of students were asked to complete a survey, and the information collected from eight of these students is shown in the following table:

Height (cm)	*Weight (kg)*	*Age (years)*	*Study mode (C on-campus, O online)*	*Fitness level (1 high, 2 medium, 3 low)*	*Pulse rate (beats/min)*
173	57	18	C	2	86
179	58	19	C	2	82
167	62	18	C	1	96
195	84	18	O	1	71
173	64	18	C	3	90
184	74	22	O	3	78
175	60	19	O	3	88
140	50	34	C	3	70

Since the answers to each of the questions in the survey will vary from student to student, each question defines a different **variable** namely:

- *height* (in centimetres)
- *weight* (in kilograms)
- *age* (in years)
- *study mode* (C = on-campus, O = online)
- *fitness level* (1 = high, 2 = medium, 3 = low)
- *pulse rate* (beats/minute).

The values we collect about each of these variables are called **data**.

Types of variables

Variables come in two general types, **categorical** and **numerical**:

Categorical variables

Categorical variables classify or name a quality or attribute– for example, a person's eye colour, study mode, or fitness level.

Data generated by a categorical variable can be used to organise individuals into one of several groups or categories that characterise this quality or attribute.

ISBN 978-1-009-11041-9

Cambridge University Press

For example, a ‘C’ in the *Study mode* column indicates that the student studies on-campus, while a ‘3’ in the *Fitness level* column indicates that their fitness level is low.

Categorical variables can be further classified as one of two types: **nominal** and **ordinal**.

- **Nominal variables**

  Nominal variables have data values that are simply names.

  The variable *Study mode* is an example of a nominal variable because the values of the variable, on-campus or online, simply name the group to which the students belong.

- **Ordinal variables**

  Ordinal variables have data values that can be used to both name and order.

  The variable *Fitness level* is an example of an ordinal variable. The data generated by this variable contains two pieces of information. First, each data value can be used to group the students by fitness level. Second, it allows us to logically order these groups according to their fitness level – in this case, as ‘low’, ‘medium’ or ‘high’.

Numerical variables

Numerical variables have data values which are quantities, generally arising from counting or measuring.

For example, a ‘179’ in the *Height* column indicates that the person is 179 cm tall, while an ‘82’ in the *Pulse rate* column indicates that they have a pulse rate of 82 beats/minute.

Numerical variables can be further classified as one of two types: **discrete** and **continuous**.

- **Discrete variables**

  Discrete variables are those which may take on only a countable number of distinct values such as 0, 1, 2, 3, 4, . . .

  Discrete random variables often arise when the situation involves counting. The number of mobile phones in a house is an example of a discrete variable.

  As a guide, discrete variables arise when we ask the question ‘How many?’

- **Continuous variables**

  Continuous variables are ones which take an infinite number of possible values, and are often associated with measuring rather than counting.

  Thus, even though we might record a person’s height as 179 cm, their height could be any value between 178.5 and 179.4 cm. We have just rounded to 179 cm for convenience, or to match the accuracy of the measuring device.

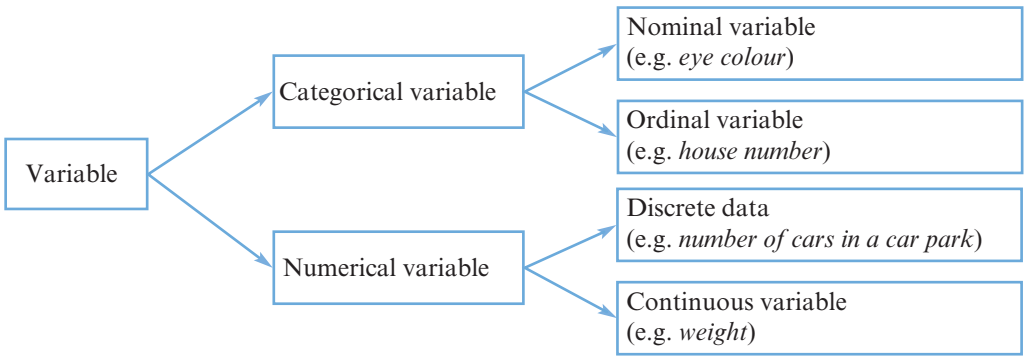
ISBN 978-1-009-11041-9

Cambridge University Press

As a guide, continuous variables arise when we ask the question 'How much?'

Comparing numerical and categorical variables

The interrelationship between categorical (nominal and ordinal) and numerical variables (discrete and continuous) is displayed in the diagram below.



Numerical or categorical?

Deciding whether data are numerical or categorical is not an entirely trivial exercise. Two things that can help your decision-making are:

1. Numerical data can always be used to perform arithmetic computations. This is not the case with categorical data. For example, it makes sense to calculate the average weight of a group of individuals, but not the average house number in a street. This is a good test to apply when in doubt.
2. It is not the variable name alone that determines whether data are numerical or categorical; it is also the way the data are recorded. For example, if the data for variable *weight* are recorded in kilograms, they are numerical. However, if the data are recorded as 'underweight', 'normal weight', 'overweight', they are categorical.

Example 1 **Types of data**

Classify the following data as nominal, ordinal, discrete or continuous.

a The number of chocolate chips in each of 10 cookies is counted.

b The time taken for 20 students to complete a puzzle is recorded in seconds.

c Member of a football club were asked to rate how they felt about the current coach, 1= Very satisfied, 2 = Satisfied, 3 = Indifferent, 4 = Dissatisfied, 5 = Very dissatisfied.

d Students are asked to each choose their preferred colour from the list 1= Blue, 2 = Green, 3 = Red, 4 = Yellow.

e Students weights were classified as 'less than 60kg', '60kg - 80kg' or 'more than 80kg'.

Solution

a **Discrete**, as the number of chocolate chips will only take whole number values.

b **Continuous**, as the data can take any value, limited only by the accuracy to which the time can be measured.

ISBN 978-1-009-11041-9

c **Ordinal**, as the numbers in this data do not represent quantities, they represent each person's level of approval of the coach.

d **Nominal**, as the numbers in this data again do not represent quantities, they represent colours.

e **Ordinal**, as each student's weight has been recorded into three categories which can be ordered.

Skill-sheet

Exercise 1A

Types of variables: categorical or numerical

Example 1

1 Classify each of the following variables (in *italics*) as categorical or numerical:

a *time* (in minutes) spent exercising each day

b *number* of frogs in a pond

c *bank account numbers*

d *height* (short, average, tall)

e *time* spent playing computer games (hours)

f *number of people* in a bus

g *eye colour* (brown, blue, green)

h *post code*

Categorical variables: nominal or ordinal

2 Classify the categorical variables identified below (in italics) as nominal or ordinal.

a The *colour* of a pencil

b The different *types of animals* in a zoo

c The *floor levels* in a building (0, 1, 2, 3 . . .)

d The *speed* of a car (less than 50 km/hr, 50 to 80 km/hr, above 80 km/hr)

e *Shoe size* (6, 8, 10, . . .)

f Family names

Numerical variables: discrete or continuous

3 Classify the numerical variables identified below (in italics) as discrete or continuous.

a The *number of pages* in a book

b The *cost* (to the nearest dollar) to fill the tank of a car with petrol

c The *volume* of petrol (in litres) used to fill the tank of a car

d The *speed* of a car in km/h

e The *number* of people at a football match

f The air *temperature* in degrees Celsius

ISBN 978-1-009-11041-9

Exam 1 style questions

4 Respondents to a survey question "Are you concerned about climate change?" were asked to select from the following responses
1 = not at all, 2 = a little, 3 = somewhat, 4 = extremely
The data which was collected in response to this question is:

A nominal **B** ordinal **C** discrete
D continuous **E** numerical

5 The variables *weight* (light, medium, heavy) and *age* (under 25 years, 25-40 years, over 40 years) are:

A a nominal and an ordinal variable respectively
B an ordinal and a nominal variable respectively
C both nominal variables
D both ordinal variables
E both continuous variables

6 Data relating to the following five variables were collected from a group of university students:

- *course*
- *study mode* (1= on campus, 2 = online)
- *study load* (1= full-time, 2-part-time)
- *number of contact hours per week*
- *number of subject needed to complete degree*

The number of these variables that are discrete numerical variables is:

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

1B Displaying and describing the distributions of categorical variables

Learning intentions

- To be able to construct frequency and percentage frequency tables for categorical data.
- To be able to construct bar charts and percentage bar charts from frequency tables.
- To be able to interpret and describe frequency tables and bar charts.
- To be able to answer statistical questions about a categorical variable.

ISBN 978-1-009-11041-9

The frequency table

With a large number of data values, it is difficult to identify any patterns or trends in the raw data.

For example, the following set of categorical data, listing the study mode (C = on-campus, O = online) for 60 individuals, is hard to make sense of.

O O C O O O C O C C C O C O O O C C C O
C O C O C C C O C O C O C O O O C O C O
C O O O O O O C C O C O O O C O C C C O

To help make sense of the data, we first need to organise them into a **frequency table**.

The frequency table

A frequency table is a listing of the values a variable takes in a data set, along with how often (frequently) each value occurs.

Frequency can be recorded as a:

- number: the number of times a value occurs, or
- percentage: the percentage of times a value occurs (**percentage frequency**):

$$\text{percentage frequency} = \frac{\text{number of times a value occurs}}{\text{total number of values}} \times 100\%$$

Example 2 Frequency table for a categorical variable

A group of 11 preschool children were asked to choose between chocolate and vanilla ice-cream (C = chocolate, V = vanilla):

C V V C C V C C C V V

Construct a frequency table (including percentage frequencies) to display the data.

Explanation

1. Set up a table as shown.
2. Count up the number of chocolate (6) and vanilla (5), and record in the Number column.
3. Add to find the total number, 11 (6 + 5).
4. Convert the frequencies into percentage frequencies. e.g. percentage chocolate= $\frac{6}{11} \times 100\% = 54.5\%$
5. Finally, total the percentages and record.

Solution

	Frequency	
flavour	Number	Percentage
chocolate	6	54.5
vanilla	5	45.5
Total	11	100.0

Note that the total should always equal the total number of observations – in this case, 11, and that the percentages should add to 100%. However, if percentages are rounded to one decimal place a total of 99.9 or 100.1 is sometimes obtained. This is due to rounding error. Totalling the count and percentages helps check on your tallying and percentaging.

ISBN 978-1-009-11041-9

The bar chart

Once categorical data have been organised into a frequency table, it is common practice to display the information graphically to help identify any features that stand out in the data.

The statistical graph we use for this purpose is the **bar chart**.

The bar chart represents the key information in a frequency table as a picture. The bar chart is specifically designed to display categorical data.

In a bar chart:

- frequency (or percentage frequency) is shown on the vertical axis
- the variable being displayed is plotted on the horizontal axis
- the height of the bar (column) gives the frequency (number or percentage)
- the bars are drawn with gaps to show that each value is a separate category
- there is one bar for each category.

Example 3 Constructing a bar chart from a frequency table

The *climate type* of 23 countries is classified as 'cold', 'mild' or 'hot'. The results are summarised in the table opposite.

a What is the level of measurement of the variable *climate type*?

b Construct a frequency bar chart to display this information.

Climate type	Frequency	
	Number	Percentage
Cold	3	13.0
Mild	14	60.9
Hot	6	26.1
Total	23	100.0

Explanation

a The data enable us to both group the countries by *climate type* and put these groups in some sort of natural order according to the 'warmth' of the different climate types. The variable is ordinal.

b **1** Label the horizontal axis with the variable name, *Climate type*. Mark the scale off into three equal intervals and label them 'Cold', 'Mild' and 'Hot'.

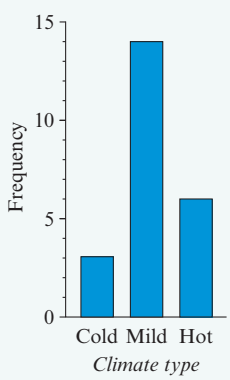
2 Label the vertical axis 'Frequency'. Scale allowing for the maximum frequency, 14.

3 For each climate type, draw a bar. There are gaps between the bars to show that the categories are separate. The height of the bar is made equal to the frequency (given in the 'Number' column).

Solution

a Ordinal

b

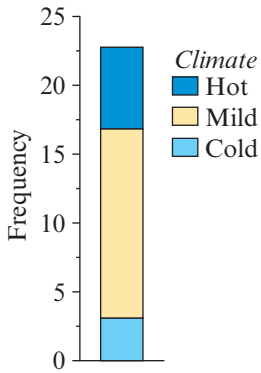


ISBN 978-1-009-11041-9

The segmented bar chart

A variation on the standard bar chart is the segmented bar chart. It is a compact display that is particularly useful when comparing two or more categorical variables.

In a **segmented bar chart**, the bars are stacked one on top of another to give a single bar with several parts or segments. The lengths of the segments are determined by the frequencies. The height of the bar gives the total frequency. A legend is required to identify which segment represents which category (see opposite). The segmented bar chart opposite was formed from the climate data used in Example 3. In a **percentage segmented bar chart**, the lengths of each segment in the bar are determined by the percentages. When this is done, the height of the bar is 100%.



Example 4 Constructing a percentage segmented bar chart from a frequency table

The climate type of 23 countries is classified as 'cold', 'mild' or 'hot'.

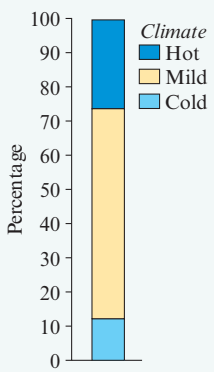
Construct a percentage frequency segmented bar chart to display this information.

Climate type	Frequency	
	Number	Percentage
Cold	3	13.0
Mild	14	60.9
Hot	6	26.1
Total	23	100.0

Explanation

1. In a segmented bar chart, the horizontal axis has no label.
2. Label the vertical axis 'Percentage'. Scale allowing for the maximum of 100 (%), Mark the scale in tens.
3. Draw a single bar of height 100. Divide the bar into three by inserting dividing lines at 13% and 73.9% (13 + 60.9%).
4. Shade (or colour) the segments differently.
5. Insert a legend to identify each shaded segment by climate type.

Solution



The mode

One of the features of a data set that is quickly revealed with a frequency table or a bar chart is the **mode** or **modal category**.

The **mode** is the most frequently occurring value or category.

ISBN 978-1-009-11041-9

In a bar chart, the mode is given by the category with the tallest bar or longest segment. For the previous bar charts, the modal category is clearly 'mild'. That is, for the countries considered, the most frequently occurring climate type is 'mild'.

Modes are particularly important in 'popularity' polls. For example, in answering questions such as 'Which is the most watched TV station between 6:00 p.m and 8:00 p.m.?' or 'When is the time a supermarket is in peak demand: morning, afternoon or night?'

Note, however, that the mode is only of real interest when a single category stands out from the others.

Answering statistical questions for categorical variables

A **statistical question** is a question that depends on data for its answer.

Statistical questions that are of most interest when working with a single categorical variable are of these forms:

- Is there a dominant category into which a large percentage of individuals fall or are the individuals relatively evenly spread across all of the categories? For example, are the shoppers in a department store predominantly male or female, or are there roughly equal numbers of males and females?
- How many and/or what percentage of individuals fall into each category? For example, what percentage of visitors to a national park are 'day-trippers' and what percentage of visitors are staying overnight?

A short written report is the standard way to answer these questions.
The following guidelines are designed to help you to produce such a report.

Guidelines for writing a report describing a categorical variable

- Briefly summarise the context in which the data were collected including the number of individuals involved in the study.
- If there is a clear modal category, ensure that it is mentioned.
- Include frequencies or percentages in the report. Percentages are preferred.
- If there are a lot of categories, it is not necessary to mention every category, but the modal category should always be mentioned.

Example 5 Describing the distribution of a categorical variable

In an investigation of the variation of climate type across countries, the climate types of 23 countries were classified as 'cold', 'mild' or 'hot'. The data are displayed in a frequency table to show the percentages.

Use the information in the frequency table to write a concise report on the distribution of climate types across these 23 countries.

	Frequency	
Climate type	Number	%
Cold	3	13.0
Mild	14	60.9
Hot	6	26.1
Total	23	100.0

ISBN 978-1-009-11041-9

Solution

Report

The climate types of 23 countries were classified as being, 'cold', 'mild' or 'hot'. The majority of the countries, 60.9%, were found to have a mild climate. Of the remaining countries, 26.1% were found to have a hot climate, while 13.0% were found to have a cold climate.

Exercise 1B

Constructing a frequency table for categorical data

Example 2

1 Construct appropriately labelled frequency tables showing both frequencies and percentage frequencies for each of the following data sets:

a *Grades*: A A C B A B B B B C C

b *Shoe size*: 8 9 9 10 8 8 8 9 8 10 12 8

Example 3

2 The following data identify the *state of residence* of a group of people where 1 = Victoria, 2 = South Australia and 3 = Western Australia.

2 1 1 1 3 1 3 1 1 3 3

a Is the variable *state of residence*, categorical or numerical?

b Construct a frequency table (with both numbers and percentages) to show the distribution of *state of residence* for this group of people.

c Construct a bar chart of the percentaged frequency table.

3 The *size* (S = small, M = medium, L = large) of 20 cars was recorded as follows.

S S L M M M L S S M

M S L S M M M S S M

a Is the variable *size* in this context numerical or categorical?

b Construct a frequency table (with both numbers and percentages) to show the distribution of size for these cars.

c Construct a percentage bar chart.

Constructing a percentage segmented bar chart from a frequency table

Example 4

4 The table shows the frequency distribution of the place of birth for 500 Australians.

Place of birth	Percentage
Australia	78.3
Overseas	21.8
Total	100.1

a Is *place of birth* an ordinal or a nominal variable?

b Display the data in the form of a percentage segmented bar chart.

ISBN 978-1-009-11041-9

5 The table records the number of new cars sold in Australia during the first quarter of one year, categorised by *type of vehicle* (private, commercial).

Type of vehicle	Frequency	
	Number	Percentage
Private	132 736	
Commercial	49 109	
Total		

a Is *type of vehicle* an ordinal or a nominal variable?

b Copy and complete the table giving the percentages correct to the nearest whole number.

c Display the data in the form of a percentage segmented bar chart.

Using the distribution of a categorical variable to answer statistical questions

Example 5

6 The table shows the frequency distribution of *school type* for a number of schools. The table is incomplete.

School type	Frequency	
	Number	Percentage
Catholic	4	20
Government	11	
Independent	5	25
Total		100

a Write down the information missing from the table.

b How many schools are categorised as 'independent'?

c How many schools are there in total?

d What percentage of schools are categorised as 'government'?

e Use the information in the frequency table to complete the following report describing the distribution of school type for these schools.

Report

_____ schools were classified according to school type. The majority of these schools, _____ %, were found to be _____. Of the remaining schools, _____ were _____ while 20% were _____.

7 Twenty-two students were asked the question, 'How often do you play sport?', with the possible responses: 'regularly', 'sometimes' or 'rarely'. The distribution of responses is summarised in the frequency table.

Plays sport	Frequency	
	Number	Percentage
Regularly	5	22.7
Sometimes	10	
Rarely		31.8
Total	22	

a Write down the information missing from the table.

b Use the information in the frequency table to complete the following report describing the distribution of student responses to the question, 'How often do you play sport?'

ISBN 978-1-009-11041-9

Report

When ____ students were asked the question, 'How often do you play sport', the dominant response was 'Sometimes', given by ____ % of the students. Of the remaining students, ____ % of the students responded that they played sport ____ while ____ % said that they played sport ____.

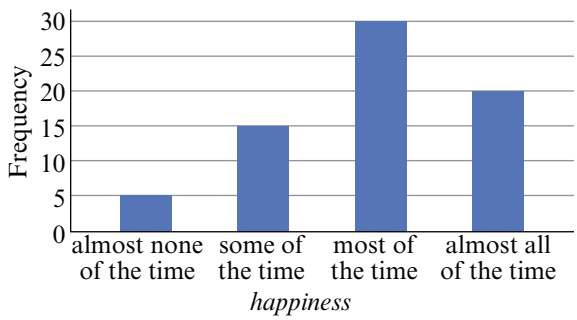
8 The table shows the frequency distribution of the eye colour of 11 preschool children.

Use the information in the table to write a brief report describing the frequency distribution of eye colour.

Eye colour	Frequency	
	Number	Percentage
Brown	6	54.5
Hazel	2	18.2
Blue	3	27.3
Total	11	100.0

Exam 1 style questions

9 In a survey people were asked to select how much of the time they felt happy from the alternatives 'almost none of the time', 'some of the time', 'most of the time', or 'almost all of the time'. Their responses are summarised in the following barchart of the variable *happiness*.



The percentage of people who chose the modal response to this question is closest to:

A 30% **B** 43% **C** 50% **D** 57% **E** 70%

1C Displaying and describing numerical data

Learning intentions

- To be able to construct frequency tables for discrete numerical data.
- To be able to construct frequency tables for grouped numerical data (discrete and continuous).
- To be able to construct a histogram from a frequency table for numerical data.
- To be able to construct a histogram from numerical data using a CAS calculator.
- To be able to describe the distribution of numerical data according to its key characteristics of shape, centre, spread and outliers.

ISBN 978-1-009-11041-9

Frequency tables can also be used to organise numerical data. For a discrete variable which only takes a small number of values the process is the same as that for categorical data, as shown in the following example.

Example 6 Constructing a frequency table for discrete numerical data taking a small number of values

The number of bedrooms in each of the 24 properties sold in a certain area over a one month period are as follows:

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

Construct a table for these data showing both frequency and percentage frequency, giving the values of the percentage frequency rounded to one decimal place.

Explanation

1. Find the maximum and the minimum values in the data set. Here the minimum is 1 and the maximum is 5.
2. Construct a table as shown, including all the values between the minimum and the maximum.
3. Count the number of 1s, 2s, etc. in the data set. Record these values in the number column and add the frequencies to find the total.
4. Convert the frequencies to percentages, and record in the per cent (%) column. For example, percentage of 1s equals $\frac{2}{24} \times 100 = 8.3\%$.
5. Total the percentages and record.

Solution

Number of bedrooms	Frequency	
	Number	%
1	2	8.3
2	5	20.8
3	7	29.2
4	8	33.3
5	2	8.3
Total	24	99.9

The grouped frequency distribution

When the variable can take on a large range of values (e.g., age from 0 to 100 years) or when the variable is continuous (e.g. response times measured in seconds to 2 decimal places), we group the data into a small number of convenient intervals.

These grouping intervals should be chosen according to the following principles:

- Every data value should be in an interval.
- The intervals should not overlap.
- There should be no gaps between the intervals.

The choice of intervals can vary but there are some guidelines.

- A division which results in about 5 to 15 groups, is preferred.

ISBN 978-1-009-11041-9

- Choose an interval width that is easy for the reader to interpret such as 10 units, 100 units or 1000 units (depending on the data).
- By convention, the beginning of the interval is given the appropriate exact value, rather than the end. As a result, intervals of 0–49, 50–99, 100–149 would be preferred over the intervals 1–50, 51–100, 101–150 etc.

When we then organise the data into a frequency table using these data intervals we call this table a **grouped frequency table**.

Example 7 **Constructing a grouped frequency table**

The data below give the average hours worked per week in 23 countries.

35.0 48.0 45.0 43.0 38.2 50.0 39.8 40.7 40.0 50.0 35.4 38.8
40.2 45.0 45.0 40.0 43.0 48.8 43.3 53.1 35.6 44.1 34.8

Construct a grouped frequency table with five intervals.

Explanation

1. Set up a table as shown. Use five intervals: 30.0–34.9, 35.0–39.9, . . . , 50.0–54.9.
2. List these intervals, in ascending order, under *Average hours worked*.
3. Count the number of countries whose average working hours fall into each of the intervals. Record these values in the 'Number' column.
4. Convert the counts into percentages and record in the 'Percentage' column.
5. Total the number and percentage columns.

Solution

Average hours worked	Frequency	
	Number	Percentage
30.0–34.9	1	4.3
35.0–39.9	6	26.1
40.0–44.9	8	34.8
45.0–49.9	5	21.7
50.0–54.9	3	13.0
Total	23	99.9

The histogram and its construction

As with categorical data, we would like to construct a visual display of a frequency table for numerical data. The graphical display of a frequency table for a numerical variable is called a **histogram**. A histogram looks similar to a bar chart but, because the data is numerical, there is a natural order to the plot and the bar widths depend on the data values.

Constructing a histogram from a frequency table

In a frequency histogram:

- frequency (count or per cent) is shown on the vertical axis
- the values of the variable being displayed are plotted on the horizontal axis

ISBN 978-1-009-11041-9

- each bar in a histogram corresponds to a data interval
- the height of the bar gives the frequency (or the percentage frequency).

Example 8 Constructing a histogram from a frequency table

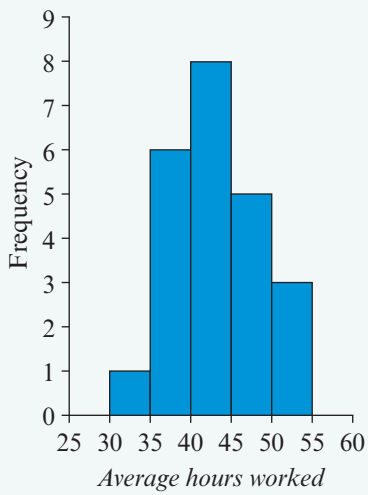
Construct a histogram for the frequency table opposite.

Average hours worked	Frequency
30.0–34.9	1
35.0–39.9	6
40.0–44.9	8
45.0–49.9	5
50.0–54.9	3
Total	23

Explanation

1. Label the horizontal axis with the variable name, *Average hours worked*. Mark the scale using the start of each interval: 30, 35, . . .
2. Label the vertical axis 'Frequency'. Scale allowing for the maximum frequency, 8.
3. Finally, for each interval draw a bar, making the height equal to the frequency.

Solution



Constructing a histogram from raw data

It is relatively quick to construct a histogram from a frequency table. However, if you have only raw data (as you mostly do), it is a very slow process because you have to construct the frequency table first. Fortunately, a CAS calculator will do this for you.

CAS 1: How to construct a histogram using the TI-Nspire CAS

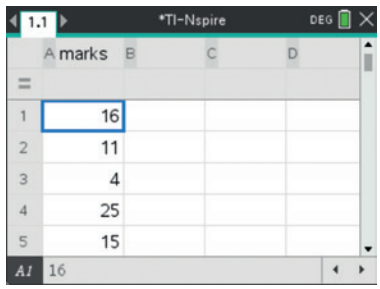
Display the following set of 27 marks in the form of a histogram.

16 11 4 25 15 7 14 13 14 12 15 13 16 14
15 12 18 22 17 18 23 15 13 17 18 22 23

ISBN 978-1-009-11041-9

Steps

1 Start a new document by pressing ctrl + N (or on>**New**. If prompted to save an existing document, move the cursor to **No** and press enter.

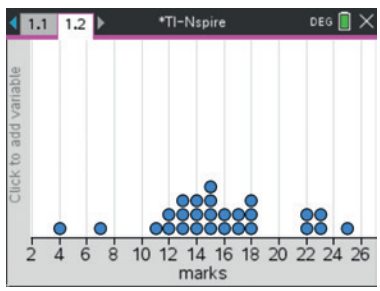


2 Select **Add Lists & Spreadsheet**.
Enter the data into a list named *marks*.

a Move the cursor to the name cell of column A and type in *marks* as the list variable. Press enter.

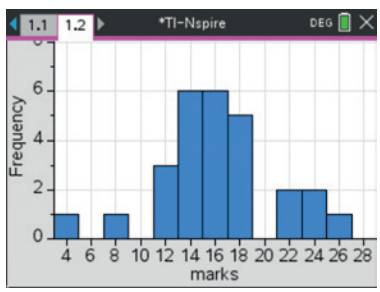
b Move the cursor down to row 1, type in the first data value and press enter. Continue until all the data have been entered. Press enter after each entry.

3 Statistical graphing is done through the **Data & Statistics** application. Press ctrl + I (or alternatively press ctrl doc) and select **Add Data & Statistics**.



a Press tab enter (or click on the **Click to add variable box** on the *x*-axis) to show the list of variables. Select *marks*.
Press enter to paste *marks* to that axis.

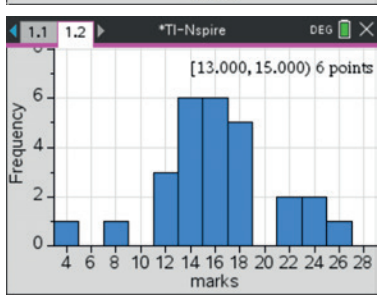
b A dot plot is displayed as the default. To change the plot to a histogram, press menu>**Plot Type>Histogram**. The histogram shown opposite has a column (or bin) width of 2, and a starting point (alignment) of 3. See Step 5 below for instructions on how to change the appearance of a histogram.



4 Data analysis

a Move the cursor over any column; a ✎ will appear and the column data will be displayed as shown opposite.

b To view other column data values, move the cursor to another column.



Note: If you click on a column, it will be selected.

Hint: If you accidentally move a column or data point, ctrl + esc enter will undo the move.

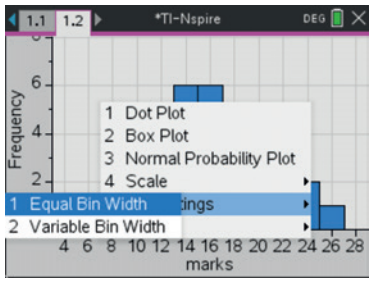
5 Change the histogram column (bin) width to **4** and the starting point to **2**.

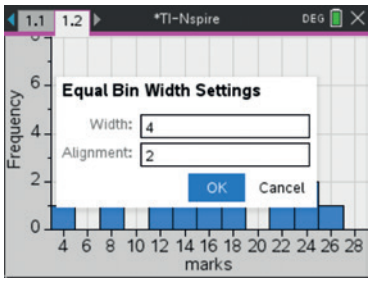
a Press ctrl + menu to access the context menu as shown (below left).
Hint: Pressing ctrl + menu enter with the cursor on the histogram gives you a context menu that relates only to histograms. You can access the commands through menu>**Plot Properties**.

b Select **Bin Settings>Equal Bin Width**.

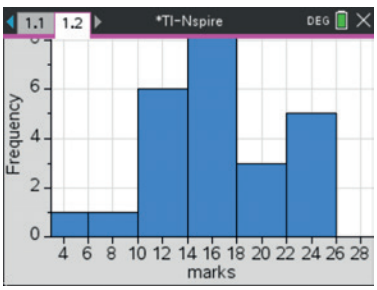
ISBN 978-1-009-11041-9

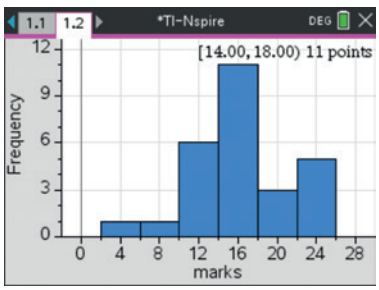
c In the settings menu (below right) change the Width to 4 and the Starting Point (Alignment) to 2 as shown. Press enter.



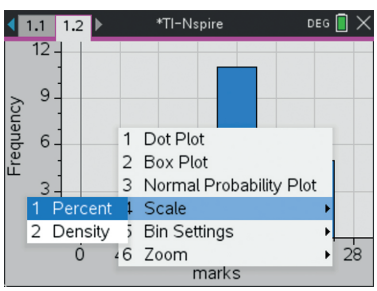


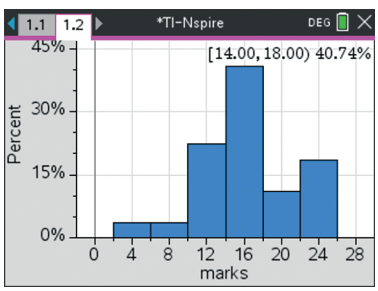
d A new histogram is displayed with column width of 4 and a starting point of 2 but it no longer fits the window (below left). To solve this problem, press ctrl + menu>**Zoom>Zoom-Data** and enter to obtain the histogram as shown below right.





6 To change the frequency axis to a percentage axis, press ctrl + menu»**Scale>Percent** and then press enter.





CAS 1: How to construct a histogram using the ClassPad

Display the following set of 27 marks in the form of a histogram.

16 11 4 25 15 7 14 13 14 12 15 13 16 14
15 12 18 22 17 18 23 15 13 17 18 22 23

ISBN 978-1-009-11041-9

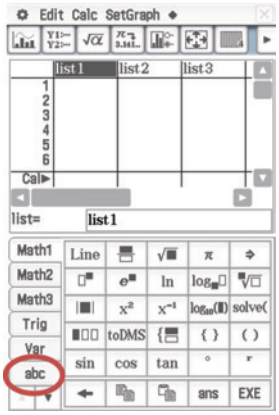
Cambridge University Press

Steps

1 From the application menu screen, locate the built-in **Statistics** application.

Tap Statistics to open.

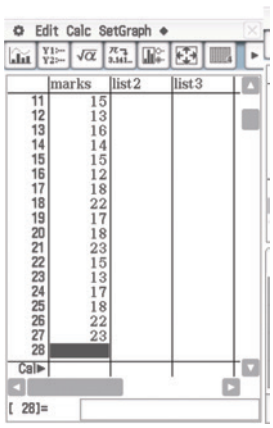
Tapping Menu from the icon panel (just below the touch screen) will display the application menu if it is not already visible.

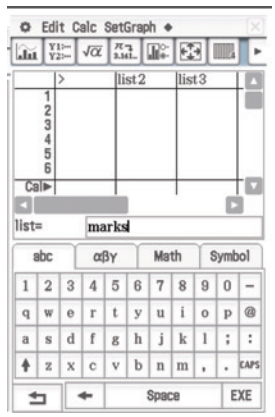


2 Enter the data into a list named *marks*.

To name the list:

- **a** Highlight the heading of the first list by tapping it.
- **b** Press Keyboard on the front of the calculator and tap the abc tab.
- **c** To enter the data, type the word **marks** and press EXE. Tap ⮌ and Keyboard to return to the list screen.
- **d** Type in each data value and press EXE or ▼ (which is found on the cursor button on the front of the calculator) to move down to the next cell.

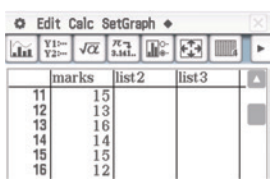


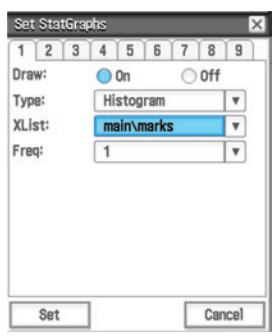


The screen should look like the one shown above right.

3 Set up the calculator to plot a statistical graph.

- **a** Tap ⊞ from the toolbar. This opens the **Set StatGraphs** dialog box.
- **b** Complete the dialog box as given below.
  - **Draw:** select **On**.
  - **Type:** select **Histogram** (▼).
  - **XList**: select **main\marks** (▼).
  - **Freq**: leave as **1**.
- **c** Tap **Set** to confirm your selections.





ISBN 978-1-009-11041-9

Note: To make sure only this graph is drawn, select **SetGraph** from the menu bar at the top and confirm that there is a tick only beside **StatGraph1** and no others.

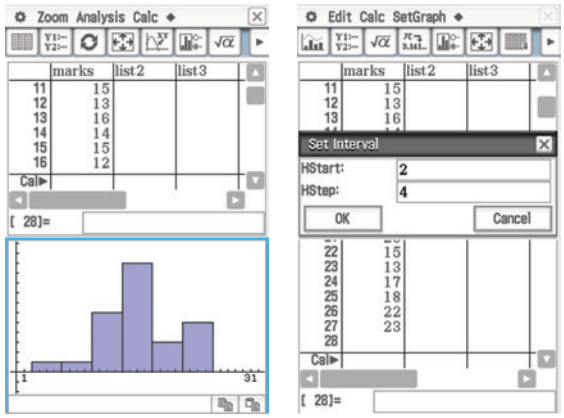
4 To plot the graph:

a Tap in the toolbar.

b Complete the **Set Interval** dialog box as follows.

- **HStart:** type **2** (i.e. the starting point of the first interval)
- **HStep:** type **4** (i.e. the interval width).

Tap OK to display histogram.



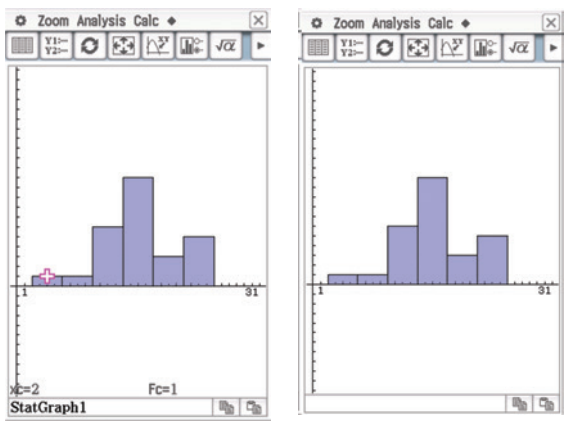
Note: The screen is split into two halves, with the graph displayed in the bottom half, as shown above. Tapping Resize from the icon panel allows the graph to fill the entire screen. Tap Resize again to return to half-screen size.

5 Tapping from the toolbar places a marker (+) at the top of the first column of the histogram (see opposite) and tells us that:

a the first interval begins at 2 **($x_c = 2$)**

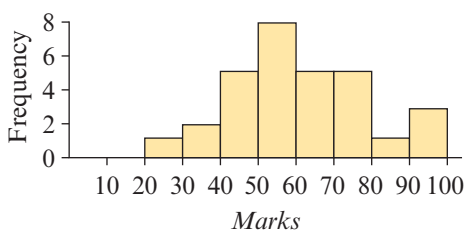
b for this interval, the frequency is 1 **($F_c = 1$)**.

To find the frequencies and starting points of the other intervals, use the cursor key arrow (▶) to move from interval to interval.



What to look for in a histogram

A histogram provides a graphical display of a data distribution. For example, the histogram opposite displays the distribution of test marks for a group of 32 students.



The purpose of constructing a histogram is to help understand the key features of the data distribution. These features are:

■ **shape** ■ **centre** ■ **spread** ■ **outliers**

ISBN 978-1-009-11041-9

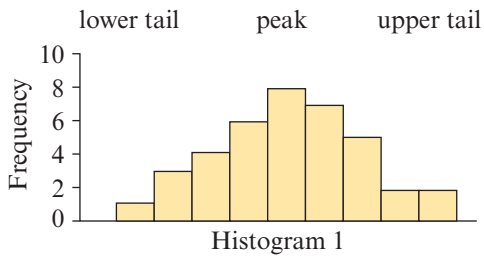
Cambridge University Press

Shape

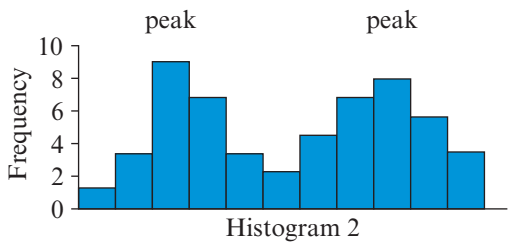
How are the data distributed? Is the histogram peaked? That is, do some data values tend to occur much more frequently than others, or is the histogram relatively flat, showing that all values in the distribution occur with approximately the same frequency?

Symmetric distributions

If a histogram is single-peaked, does the histogram region tail off evenly on either side of the peak? If so, the distribution is said to be **symmetric** (see Histogram 1).



Histogram 1



Histogram 2

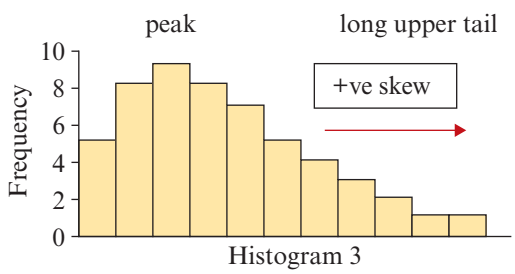
A single-peaked **symmetric distribution** is characteristic of the data that derive from measuring variables such as intelligence test scores, weights of oranges, or any other data for which the values vary evenly around some central value.

The double-peaked distribution (histogram 2) is symmetric about the dip between the two peaks. A histogram that has two distinct peaks indicates a **bimodal** (two modes) distribution.

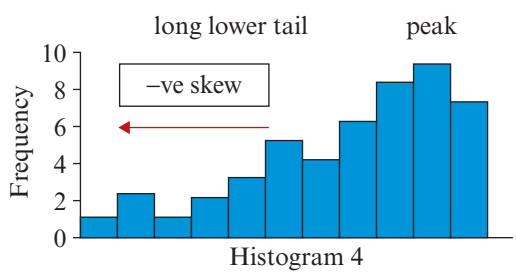
A bimodal distribution often indicates that the data have come from two different populations. For example, if we were studying the distance the discus is thrown by Olympic-level discus throwers, we would expect a bimodal distribution if both male and female throwers were included in the study.

Skewed distributions

Sometimes a histogram tails off primarily in one direction. If a histogram tails off to the right, we say that it is **positively** skewed (Histogram 3). The distribution of salaries of workers in a large organisation tends to be positively skewed. Most workers earn a similar salary with some variation above or below this amount, but a few earn more and even fewer, such as the senior manager, earn even more. The distribution of house prices also tends to be positively skewed.



Histogram 3



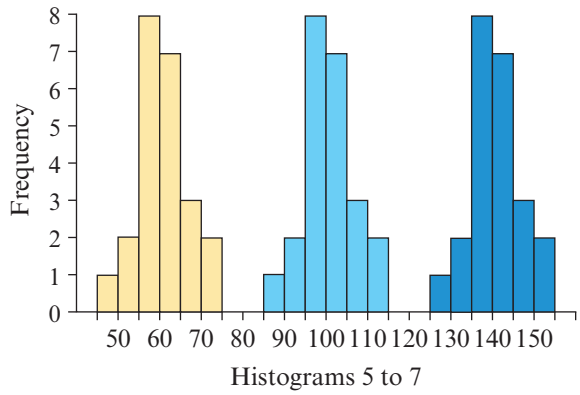
Histogram 4

If a histogram tails off to the left, we say that it is **negatively** skewed (Histogram 4). The distribution of age at death tends to be negatively skewed. Most people die in old age, a few in middle age and fewer still in childhood.

ISBN 978-1-009-11041-9

Centre

Histograms 6 to 8 display the distribution of test scores for three different classes taking the same subject. They are identical in shape, but differ in where they are located along the axis. In statistical terms we say that the distributions are 'centred' at different points along the axis. But what do we mean by the **centre of a distribution**?



Histograms 5 to 7

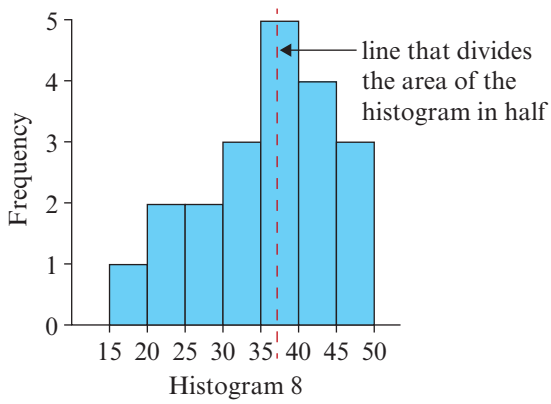
This is an issue we will return to in more detail later in the chapter. For the present we will take the centre to be the middle of the distribution.

The middle of a symmetric distribution is reasonably easy to locate by eye. Looking at histograms 5 to 7, it would be reasonable to say that the centre or middle of each distribution lies roughly halfway between the extremes; half the observations would lie above this point and half below. Thus we might estimate that histogram 5 (yellow) is centred at about 60, histogram 6 (light blue) at about 100, and histogram 7 (dark blue) at about 140.

For skewed distributions, it is more difficult to estimate the middle of a distribution by eye. The middle is not halfway between the extremes because, in a skewed distribution, the scores tend to bunch up at one end.

However, if we imagine a cardboard cut-out of the histogram, then the middle lies on the line that divides the histogram into two equal areas (Histogram 8).

Using this method, we would estimate the centre of the distribution to lie somewhere between 35 and 40, but closer to 35, so we might opt for 37. However, remember that this is only an estimate.



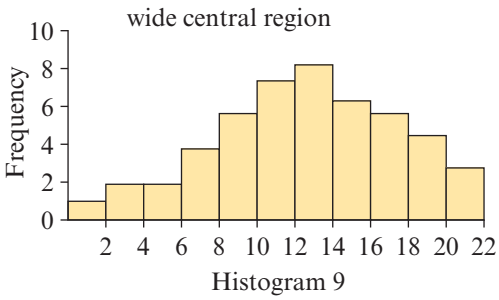
Histogram 8

Spread

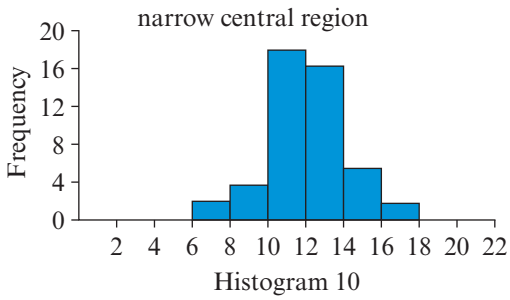
If the histogram is single-peaked, is it narrow? This would indicate that most of the data values in the distribution are tightly clustered in a small region. Or is the peak broad? This would indicate that the data values are more widely spread out. Histograms 9 and 10 are both single-peaked. Histogram 9 has a broad peak, indicating that the data values are not very tightly clustered about the centre of the distribution. In contrast, Histogram 10 has a narrow peak, indicating that the data values are tightly clustered around the centre of the distribution.

ISBN 978-1-009-11041-9

Cambridge University Press



Histogram 9

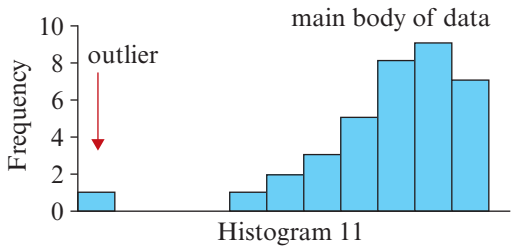


Histogram 10

Outliers

Outliers are any data values that stand out from the main body of data. These are data values that are atypically high or low. See, for example, Histogram 11, which shows an outlier. In this case it is a data value that is atypically low compared to the rest of the data values.

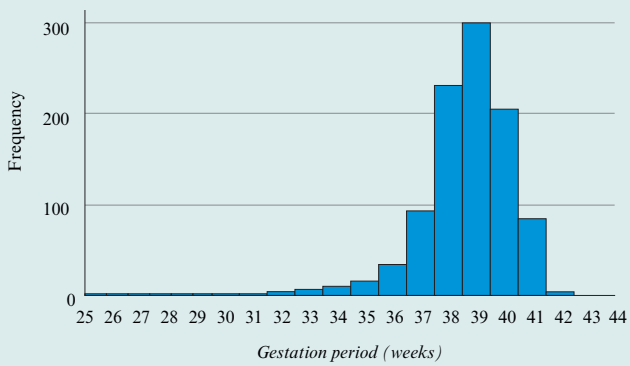
In the histogram shown there appears to be an outlier, a data value which is lower than than rest of the data values. Such values should be checked, they may indicate an unusually low value, but they may also indicate an error in the data.



Histogram 11

Example 9 Describing the features of a distribution from a histogram

The histogram shows the gestation period (completed weeks) for a sample for 1000 babies born in Australia one year. Describe this histogram in terms of shape, centre, spread and outliers.



Explanation

1 Determine the shape of the distribution.

2 Locate the (approximate) centre of the distribution, the value seems to divide the area of the histogram in half.

Solution

1 The distribution is clearly negatively skewed, with a long lower tail.

2 The centre of the distribution is around 38-39 weeks.

ISBN 978-1-009-11041-9

Cambridge University Press

3 Consider the spread of the distribution, are the majority of the values close to the centre, or quite spread out?	**3** The data values range from 25-42 weeks, but most of the data values are close to the centre, in the range of 36-41 weeks.
4 Can we identify any outliers?	**4** Those values which are less than 34 weeks seem to be small in comparison to the rest of the data, but which values are outliers cannot be determined from the histogram.

We can see from this example that it is very difficult to give an exact values for centre and spread, and to clearly identify outliers, from the histogram. We will return to this example later in the chapter once we have discussed which measures of centre and spread are appropriate for this distribution, and when we have an exact definition on an outlier.

Exercise 1C

Constructing a frequency table for discrete numerical data taking a small number of values

Example 6

1 The number times a sample of 20 people bought take-away food over a one week period is as follows:

0 5 3 0 1 0 2 4 3 1 0 2 1 2 1 5 3 0 0 4

a Construct a frequency table for the data, including both the frequency and percentage frequency.

b What percentage of people bought take away food more than 3 times?

c What is the mode of this distribution?

2 The number of chocolate chips per biscuit in a sample 40 biscuits was found to be as as follows:

2 5 4 4 5 4 6 4 4 4 5 4 4 5 6 6 5 5 4 6
4 5 5 4 5 4 6 4 6 4 5 4 5 4 6 5 5 6 4 6

a Construct a frequency table for the data, including both the frequency and percentage frequency.

b What percentage of biscuits contained three or less chocolate chips?

c What is the mode of this distribution?

Constructing a grouped frequency table

Example 7

3 The following are the heights of the 25 players in a women's football team, in centimetres.

ISBN 978-1-009-11041-9

Cambridge University Press

188	175	176	161	183
169	171	177	165	166
162	170	174	168	178
169	181	173	164	179
163	170	164	175	182

Height (cm)	Frequency
160–164	
165–169	
170–174	
175–179	
180–184	
185–189	
Total	25

a Use the data to complete the grouped frequency table.

b What is the model height for this group of players?

c What percentage of the players are 180 cm or more in height?

Constructing a histogram from a frequency table

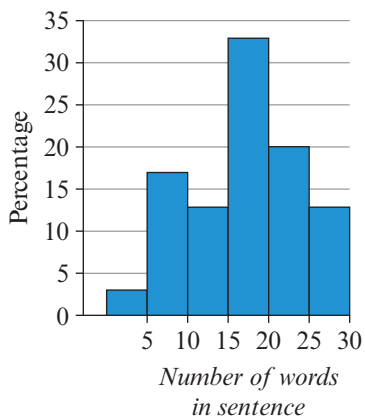
Example 8

4 Construct a histogram to display the information in the frequency table opposite. Label axes and mark scales.

Population density	Frequency
0–199	11
200–399	4
400–599	4
600–799	2
800–999	1
Total	22

Reading information from a histogram

5 The histogram opposite displays the distribution of the number of words in 30 randomly selected sentences.



a What percentage of these sentences contained:

i 5–9 words?

ii 25–29 words?

iii 10–19 words?

iv fewer than 15 words?

Write answers correct to the nearest per cent.

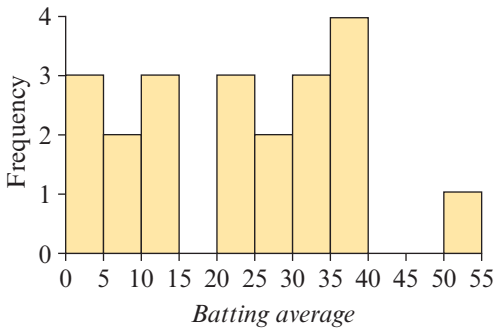
b How many of these sentences contained:

i 20–24 words? **ii** more than 25 words?

c What is the modal interval?

ISBN 978-1-009-11041-9

6 The histogram opposite displays the distribution of the average batting averages of cricketers playing for a district team.



a How many players have their averages recorded in this histogram?

b How many of these cricketers had a batting average:

i 20 or more?

ii less than 15?

iii at least 20 but less than 30?

iv of 45?

c What percentage of these cricketers had a batting average:

i 50 or more? **ii** at least 20 but less than 40?

Constructing a histogram from raw data using a CAS calculator

7 The pulse rates of 23 students are given below.

86 82 96 71 90 78 68 71 68 88 76 74
70 78 69 77 64 80 83 78 88 70 86

a Use a CAS calculator to construct a histogram so that the first column starts at 63 and the column width is two.

b **i** What is the starting point of the third column?

ii What is the 'count' for the third column? What are the *actual* data values?

c Redraw the histogram so that the column width is five and the first column starts at 60.

d For this histogram, what is the count in the interval '65 to <70'?

8 The numbers of children in the families of 25 VCE students are listed below.

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

a Use a CAS calculator to construct a histogram so that the column width is one and the first column starts at 0.5.

b What is the starting point for the fourth column and what is the count?

c Redraw the histogram so that the column width is two and the first column starts at 0.

d **i** What is the count in the interval from 6 to less than 8?

ii What actual data value(s) does this interval include?

ISBN 978-1-009-11041-9

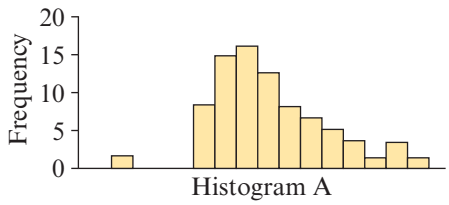
Determining shape, centre and spread from a histogram

Example 9

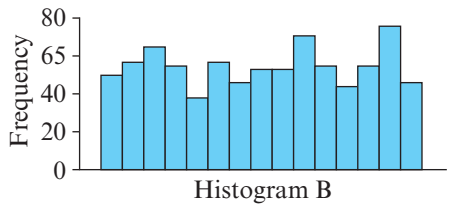
9 Identify each of the following histograms as approximately symmetric, positively skewed or negatively skewed, and mark the following.

- **i** The mode (if there is a clear mode)
- **ii** Any potential outliers
- **iii** The approximate location of the centre

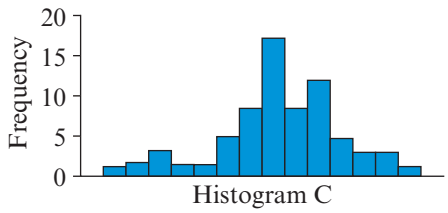
a



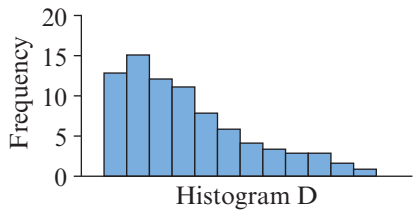
b



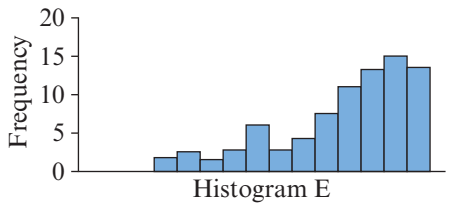
c



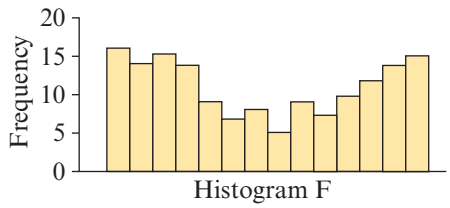
d



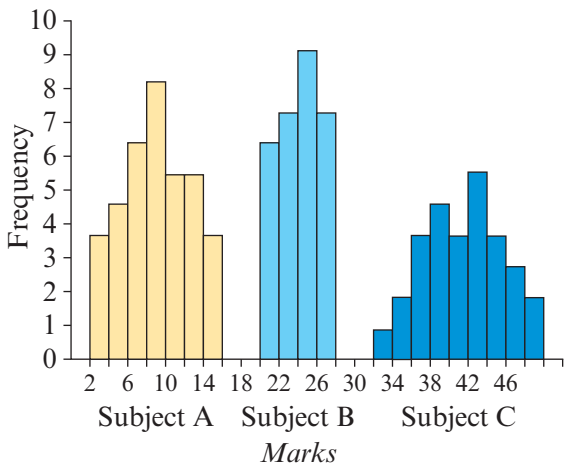
e



f



10 These three histograms show the marks obtained by a group of students in three subjects.



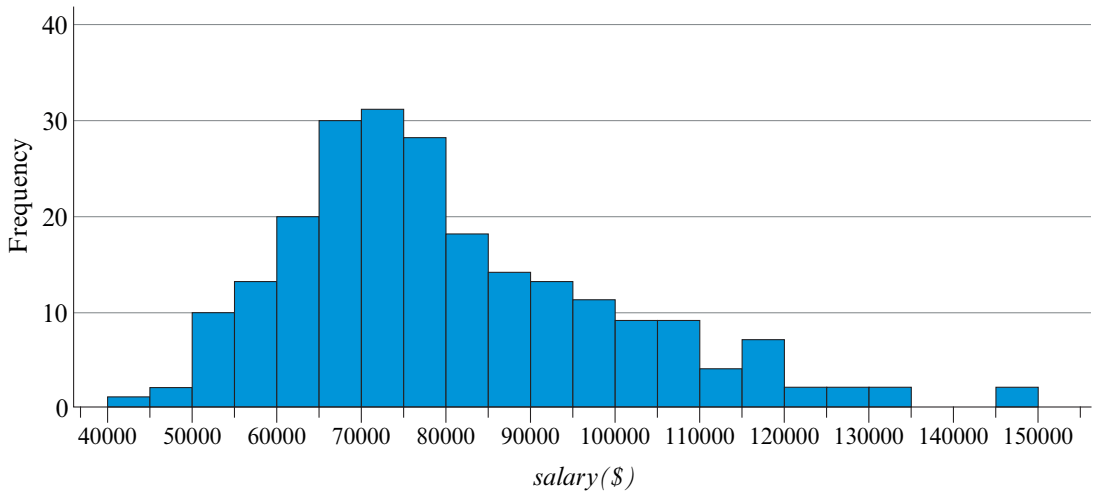
- **a** Are each of the distributions approximately symmetric or skewed?
- **b** Are there any clear outliers?
- **c** Determine the interval containing the central mark for each of the three subjects.
- **d** In which subject was the spread of marks the least?

ISBN 978-1-009-11041-9

Exam 1 style questions

Use the following information to answer questions 11 and 12

The annual salaries for all the sales staff in a large company are summarised in the following histogram.



11 The number of people in the company who earn from $65,000 to less than $70,000 per year is equal to:

A 20 **B** 30 **C** 50 **D** 32 **E** 62

12 The shape of this histogram is best described as:

A positively skewed with possible outlliers
B positively skewed with no outliers
C approximately symmetric
D negatively skewed with no outliers
E negatively skewed with outliers

1D Dot plots and stem plots

Learning intentions

- To be able to construct a dot plot for numerical data.
- To be able to construct a stem plot for numerical data.

Dot plots and stem plots are two simple plots used to display numerical data. They are generally constructed by hand (that is, without using a calculator), from a data set that is reasonably small.

The dot plot

The simplest way to display numerical data is to construct a **dot plot**. A dot plot is particularly suitable for displaying discrete numerical data and provides a very quick way to order and display a small data set.

ISBN 978-1-009-11041-9

A dot plot consists of a number line with each data point marked by a dot. When several data points have the same value, the points are stacked on top of each other.

Example 10 **Constructing a dot plot**

The ages (in years) of the 13 members of a cricket team are:

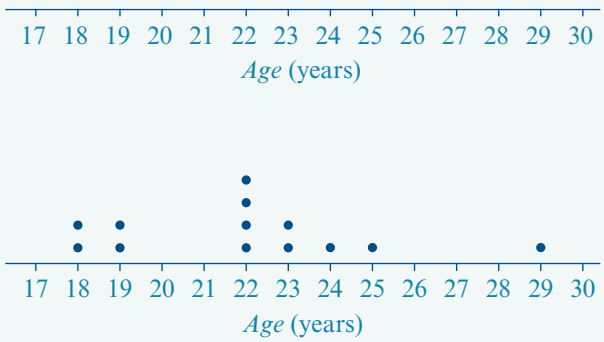
22 19 18 19 23 25 22 29 18 22 23 24 22

Construct a dot plot.

Explanation

1. Draw a number line, scaled to include all data values. Label the line with the variable being displayed.
2. Plot each data value by marking a dot above the corresponding value on the number line.

Solution



While some CAS calculators will construct a stem plot, they were designed to be a quick and easy way of ordering and displaying a small data set by hand.

The stem plot

The **stem-and-leaf plot**, or **stem plot** for short, is another quick and easy way to display numerical data. Stem plots work well for both discrete and continuous data. They are particularly useful for displaying small- to medium-sized sets of data (up to about 50 data values). Like the dot plot, they are designed to be a pen and paper technique.

In a stem plot, each data value is separated into two parts: its leading digits, which make up the 'stem' of the number, and its last digit, which is called the 'leaf'.

For example, in the stem-and-leaf plot opposite, the data values 21 and 34 are displayed as follows:

	Stem	Leaf
21 is displayed as	2	1
34 is displayed as	3	4

A key should always be included to show how the numbers in the plot should be interpreted.

Key: 1|2 = 12

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

ISBN 978-1-009-11041-9

Cambridge University Press

Example 11 Constructing a stem plot

University participation rates (%) in 23 countries are given below.

26	3	12	20	36	1	25	26	13	9	26	27
15	21	7	8	22	3	37	17	55	30	1	

Display the data in the form of a stem plot.

Explanation

1 The data set has values in the units, tens, twenties, thirties, forties and fifties. Write the stems 0, 1, 2, 3, 4, and 5 in ascending order, followed by a vertical line. Now attach the leaves. The first data value is '26'. The stem is '2' and the leaf is '6'. Opposite the 2 in the stem, write down the number 6, as shown.

2 Continue systematically working through the data, following the same procedure until all points have been plotted. You will then have the stem plot, as shown.

3 To complete the task, write the leaves on each stem in ascending order, then add the variable name and a key.

Solution

0	
1	
2	6
3	
4	
5	

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

rate (%) Key: 1|2 = 12

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

Stem plots with split stems

In some instances, using the simple process outlined above produces a stem plot that is too cramped to give an good overall picture of the variation in the data. This happens when the data values all have the same one or two first digits.

ISBN 978-1-009-11041-9

Cambridge University Press

For example, consider the marks obtained by 17 VCE students on a statistics test.

2 12 13 9 18 17 7 16 12 10 16 14 11 15 16 15 17

If we use the process described in Example 11 to form a stem plot, we end up with a 'bunched-up' plot like the below.

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

We can solve this problem by 'splitting' the stems.

Generally the stem is split into halves or fifths as shown below.

Key: 1|6 = 16

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

Single stem

Key: 1|6 = 16

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

Stem split into halves

Key: 1|6 = 16

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

Stem split into fifths

Splitting the stems is useful when there are only a few different values for the stem.

Exercise 1D

Constructing a dot plot

Example 10

1 The following data gives the number of rooms in 11 houses.

4 6 7 7 8 4 4 8 8 7 8

a Is the variable *number of rooms* discrete or continuous?

b Construct a dot plot.

2 The following data give the number of children in the families of 14 VCE students:

1 6 2 5 5 3 4 4 2 7 3 4 3 4

a Is the variable *number of children* discrete or continuous?

b Construct a dot plot.

c Write down the value of the mode. What does the mode represent in the context of the data?

3 The following data give the average life expectancies in years of 13 countries.

76 75 74 74 73 73 75 71 72 75 75 78 72

ISBN 978-1-009-11041-9

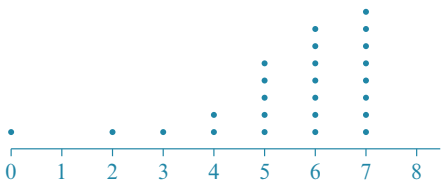
a Is the variable *life expectancy* discrete or continuous?

b Construct a dot plot.

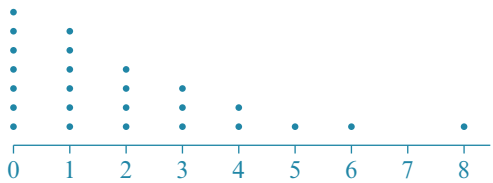
c Write down the value of the mode. What does the mode represent in the context of the data?

4 Describe the shape of each of the following distributions (negatively skewed, positively skewed, or approximately symmetric).

a



b



5 The ages of each member of a running club are as follows:

22 20 20 23 21 22 21 25 21 24 18
20 19 22 23 25 19 21 20 21 21 22

a Construct a dot plot of the ages of the players.

b What is the mode of this distribution?

c What is the shape of the distribution of runners ages?

d What percentage of runners are younger than 20? Give your answer to the nearest whole percentage.

Constructing a stem plot

Example 11

6 The data below give the urbanisation rates (%) in 23 countries.

54 99 22 20 31 3 22 9 25 3 56 12
16 9 29 6 28 100 17 99 35 27 12

a Is the variable *urbanisation rate* discrete or continuous?

b Construct a stem plot with an appropriate key.

Constructing a stem plot with split stems

7 The data below give the wrist circumference (in cm) of 15 men.

16.9 17.3 19.3 18.5 18.2 18.4 19.9 16.7
17.7 16.5 17.0 17.2 17.6 17.1 17.6

a Is the variable *wrist circumference* discrete or continuous?

b Construct a stem plot for wrist circumference using:

i stems: 16, 17, 18, 19

ii these stems split into halves, that is: 16, 16, 17, 17, . . .

Interpreting a stem plot

8 Describe the shape of each of following distributions (negatively skewed, positively skewed, or approximately symmetric).

ISBN 978-1-009-11041-9

a key: 4|1 represents 4.1

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

b key: 3|1 represents 31

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

9 The stem plot on the right shows the ages, in years, of all the people attending a meeting.

a How many people attended the meeting?

b What is the shape of the distribution of ages?

c How many of these people were less than 33 years old?

Age (years) key: 2|0 represents 20

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

Exam 1 style questions

Use the following information to answer questions 10 and 11

The following stem plot shows the distribution of the time it took (in minutes) for each of a group of 25 people to solve a complex task.

Time (minutes) key: 4|0 represents 4.0

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

10 The shape of this distribution is best described as:

A positively skewed with a possible outlier

B positively skewed with no outliers

C approximately symmetric

D negatively skewed with no outliers

E negatively skewed with outliers

ISBN 978-1-009-11041-9

11 The time taken by the slowest 20% of people was:

A more than 8.5 minutes **B** 8.5 minutes or more **C** less than 5.6 minutes
D 5.6 minutes or less **E** more than 5.6 minutes

1E Using a logarithmic (base 10) scale to display data

Learning intentions

- To be able to revise the concept of $\log_{10} x$.
- To be able to investigate the effect of the logarithmic scale on the features of a distribution.

Many numerical variables that we deal with in statistics have values that range over several orders of magnitude. For example, the populations of countries range from a few thousand to hundreds of thousands, to millions, to hundreds of millions to just over 1 billion. Constructing a histogram that effectively locates every country on the plot is impossible.

One way to solve this problem is to use a scale that spreads out the countries with small populations and 'pulls in' the countries with huge populations.

A scale that will do this is called a logarithmic scale (or, more commonly, a **log scale**).

Consider the numbers:

0.01, 0.1, 1, 10, 100, 1000, 10 000, 100 000, 1 000 000

Such numbers can be written more compactly as:

10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}

In fact, if we make it clear we are only talking about powers of 10, we can merely write down the powers:

−2, −1, 0, 1, 2, 3, 4, 5, 6

These powers are called the **logarithms** of the numbers or 'logs' for short.

When we use logarithms to write numbers as powers of 10, we say we are working with logarithms to the base 10. We can indicate this by writing $\log_{10}$.

$\log_{10} x$

If $\log_{10} x = b$, then $10^{b} = x$

Thus we can say for example that:

- $\log_{10}(100) = 2$, since $10^{2} = 100$
- $\log_{10}(1000) = 3$, since $10^{3} = 1000$
- $\log_{10}(1000000) = 6$, since $10^{6} = 1000000$

ISBN 978-1-009-11041-9

Properties of logarithms to the base 10

1 If a number is greater than one, its log to the base 10 is greater than zero.
2 If a number is greater than zero but less than one, its log to the base 10 is negative.
3 If the number is zero, then its log is undefined.

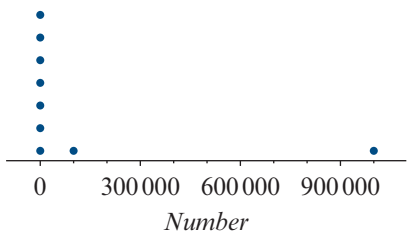
The effect of the the logarithmic scale

The set of numbers

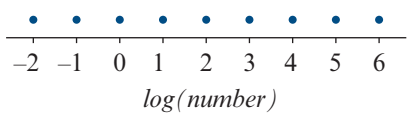
0.01, 0.1, 1, 10, 100, 1000, 10 000, 100 000, 1 000 000

ranges from 0.01 to 1 million.

Thus, if we wanted to plot these numbers on a scale, the first seven numbers would cluster together at one end of the scale, while the eighth (1 million) would be located at the far end of the scale.



By contrast, if we plot the logs of these numbers, they are evenly spread along the scale. We use this idea to display a set of data whose values range over several orders of magnitude. Rather than plot the data values themselves, we plot the logarithms of their data values.



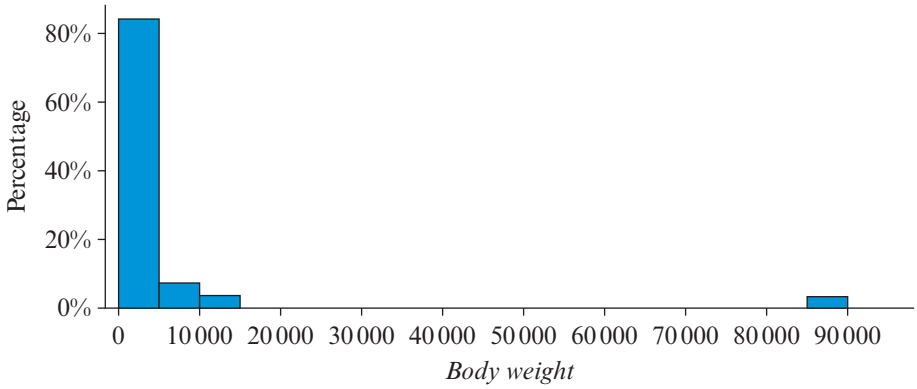
Logarithmic transformation

A **logarithmic transformation** involves changing the scale on the horizontal axis from x to $\log_{10}(x)$, and replacing each of the data values with its logarithm.

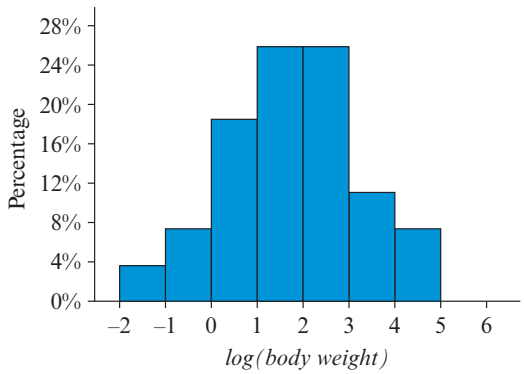
For example, the histogram below displays the body weights (in kg) of a number of animal species. Because the animals represented in this data set have weights ranging from around 1 kg to 90 tonnes (a dinosaur), most of the data are bunched up at one end of the scale and much detail is missing. The distribution of weights is highly positively skewed, with an outlier.

ISBN 978-1-009-11041-9

Cambridge University Press



However, when a logarithmic transformation is used, their weights are much more evenly spread along the scale. The distribution is now approximately symmetric, with no outliers, and the histogram is considerably more informative.



We can now see that the percentage of animals with weights between 10 and 100 kg is similar to the percentage of animals with weights between 100 and 1000 kg.

Working with logarithms

To construct and interpret a log data plot, like the one above, you need to be able to:

1. Work out the log for any number. So far we have only done this for numbers such as 10, 100, 1000 that are exact powers of 10; for example, $100 = 10^2$, so $\log 100 = 2$.
2. Work backwards from a log to the number it represents. This is easy to do in your head for logs that are exact powers of 10 – for example, if the log of a number is 3 then the number is $10^3 = 1000$. But it is not a sensible approach for numbers that are not exact powers of 10.

Your CAS calculator is the key to completing both of these tasks in practice.

Example 12 Using a CAS calculator to find logs

a Find the log of 45, correct to two significant figures.

b Find the number with log equal to 2.7125, correct to the nearest whole number.

ISBN 978-1-009-11041-9

Cambridge University Press

Explanation

a Open a calculator screen, type log (45) and press enter. Write down the answer correct to two significant figures.

b If the log of a number is 2.7125, then the number is $10^{2.7125}$.

Enter the expression $10^{2.7125}$ and press enter. Write down the answer correct to the nearest whole number.

Solution

a $\log 45 = 1.65\ldots$

$= 1.7$ (to 2 sig. figs)

b $10^{2.7125} = 515.82\ldots$

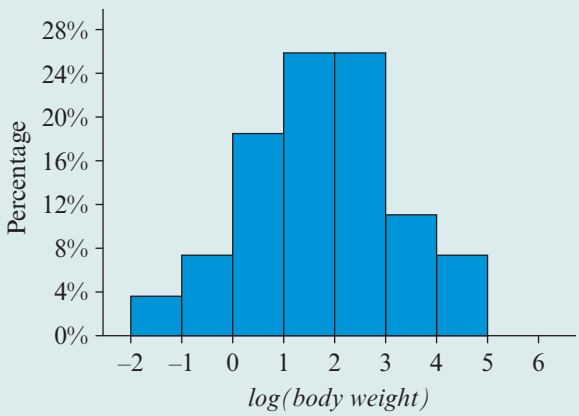
$= 516$ (to the nearest whole number)

Analysing data displays with a logarithmic scale

Now that you know how to work out the log of any number and convert logs back to numbers, you can analyse a data plot using a log scale.

Example 13 Interpreting a histogram with a log scale

The histogram shows the distribution of the weights of 27 animal species plotted on a log scale.



a What body weight (in kg) is represented by the number 4 on the log scale?

b How many of these animals have body weights more than 10 000 kg?

c The weight of a cat is 3.3 kg. Use your calculator to determine the log of its weight correct to two significant figures.

d Determine the weight (in kg) of the animal with a log(*body weight*) of 3.4 (the elephant). Write your answer correct to the nearest whole number.

Explanation

a If the log of a number is 4 then the number is $10^4 = 10\,000$.

Solution

a $10^4 = 10\,000$ kg

ISBN 978-1-009-11041-9

b On the log scale, 10 000 is shown as 4.

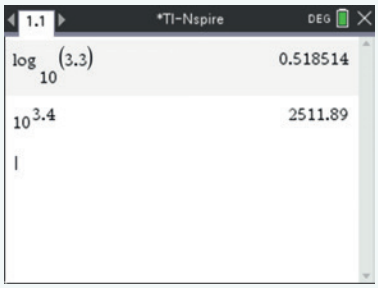
Thus, the number of animals with a weight greater than 10 000 kg corresponds to the number of animals with a log weight of greater than 4.

This can be determined from the histogram which shows there are two animals with log weights greater than 4.

c The weight of a cat is 3.3 kg. Use your calculator to find log 3.3. Write the answer correct to two significant figures.

d The log weight of an elephant is 3.4. Determine its weight in kg by using your calculator to evaluate $10^{3.4}$. Write the answer correct to the nearest whole number.

b Two animals



c Cat: $\log 3.3 = 0.518...$

$= 0.52$ kg (to 2 sig. figs)

d Elephant: $10^{3.4} = 2511.88...$

$= 2512$ kg

Constructing a histogram with a log scale

The task of constructing a histogram is also a CAS calculator task.

CAS 2: Using a TI-Nspire CAS to construct a histogram with a log scale

The weights of 27 animal species (in kg) are recorded below.

1.4	470	36	28	1.0	12 000	2600	190	520
10	3.3	530	210	62	6700	9400	6.8	35
0.12	0.023	2.5	56	100	52	87 000	0.12	190

Construct a histogram to display the distribution:

a of the body weights of these 27 animals and describe its shape

b of the log of the body weights of these animals and describe its shape.

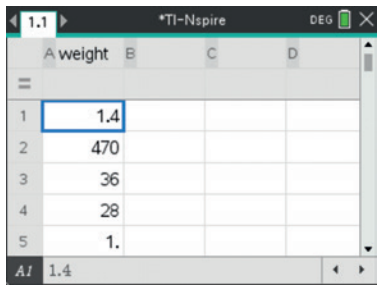
ISBN 978-1-009-11041-9

Cambridge University Press

Steps

1 a Start a new document by pressing ctrl + N.

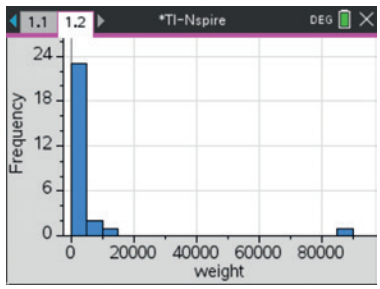
b Select **Add Lists & Spreadsheet**.
Enter the data into a column named *weight*.



2 a Press ctrl + I and select **Add Data & Statistics**.
Click on the **Click to add variable** on the *x*-axis and select the variable *weight*. A dot plot is displayed.

b Plot a histogram using menu>**Plot Type>Histogram**.

c Describe the shape of the distribution.

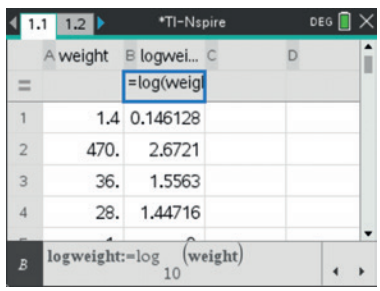


Shape: positively skewed with outliers

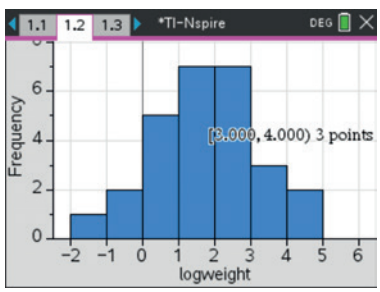
3 a Return to the **Lists & Spreadsheet** screen.

b Name another list *logweight*.

c Move the cursor to the formula cell below the *logweight* heading. Type in = **log(*weight*)**.
Press enter to calculate the values of *logweight*.



4 a Plot a histogram using a log scale. That is, plot the variable *logweight*.
Note: Use menu>**Plot Properties**>**Histogram Properties**>**Bin Settings**>**Equal Bin Width** and set the column width (bin) to 1 and alignment (start point) to −2 and use menu>**Window/Zoom**>**Zoom-Data** to rescale.



b Describe the shape of the distribution.

Shape: approximately symmetric

ISBN 978-1-009-11041-9

CAS 2: Using a ClassPad to construct a histogram with a log scale

The weights of 27 animal species (in kg) are recorded below.

1.4	470	36	28	1.0	12 000	2600	190	520
10	3.3	530	210	62	6700	9400	6.8	35
0.12	0.023	2.5	56	100	52	87 000	0.12	190

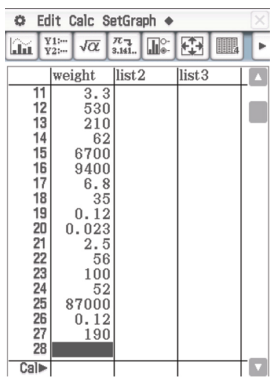
Construct a histogram to display the distribution:

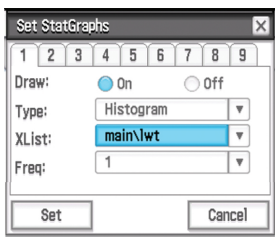
a of the body weights of these 27 animals and describe its shape

b of the log of the body weights of these animals and describe its shape.

Steps

1 In the statistics application Statistics enter the data into a column named *weight* as shown.





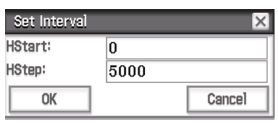
2 Plot a histogram of the data.

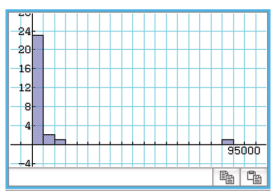
a Tap from the toolbar.

b Complete the dialog box.

- **Draw:** select **On**.
- **Type:** select **Histogram** (▼)
- **XList:** select **main\weight**(▼).
- **Freq:** leave as **1**.

Tap **Set** to confirm your selections.





c Tap in the toolbar.

d Complete the **Set Interval** dialog box as follows:

HStart: 0

HStep: 5000

Describe the shape of the distribution. Shape: positively skewed with outliers

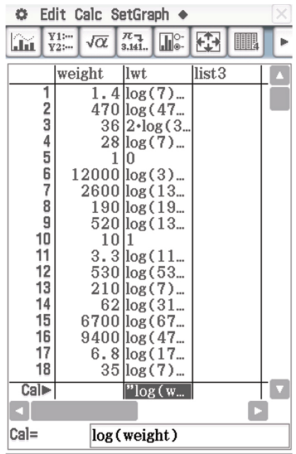
ISBN 978-1-009-11041-9

3 **a** Return to the data entry screen.

b Name another column 'lwt', short for log(*weight*).

c Tap in the calculation cell at the bottom of this column.

Type log(*weight*) and tap EXE.



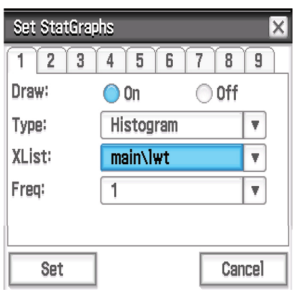
4 Plot a histogram to display the distribution of weights on a log scale. That is, plot the variable lwt.

a Tap from the toolbar.

b Complete the dialog box.

- **Draw**: select **On**.
- **Type**: select **Histogram** (▼).
- **XList**: select **main\lwt** (▼).
- **Freq**: leave as **1**.

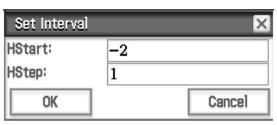
Tap **Set** to confirm your selections.



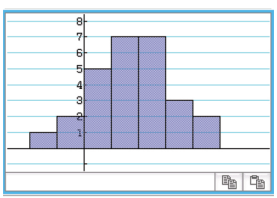
c Tap in the toolbar.

d Complete the **Set Interval** dialog box as follows:

- **HStart**: **type -2**
- **HStep**: type **1**



Tap **OK** to display histogram.



Describe the shape of the distribution.

Shape: approximately symmetric

Skill-sheet

Exercise 1E

Determining the log of a number

Example 12

1 Using a CAS calculator, find the logs of the following numbers correct to one decimal place.

a 2.5	**b** 25	**c** 250	**d** 2500
e 0.5	**f** 0.05	**g** 0.005	**h** 0.0005

ISBN 978-1-009-11041-9

Determining a number from its log

2 Find the numbers with the following logs:

a −2.5 **b** −1.5 **c** −0.5 **d** 0

Write your decimal answers correct to two significant figures.

Constructing a histogram with a log scale

3 The brain weights of the same 27 animal species (in g) are recorded below.

465	423	120	115	5.5	50	4600	419	655
115	26	680	406	1320	5712	70	179	56
1.0	0.4	12	175	157	440	155	3.0	180

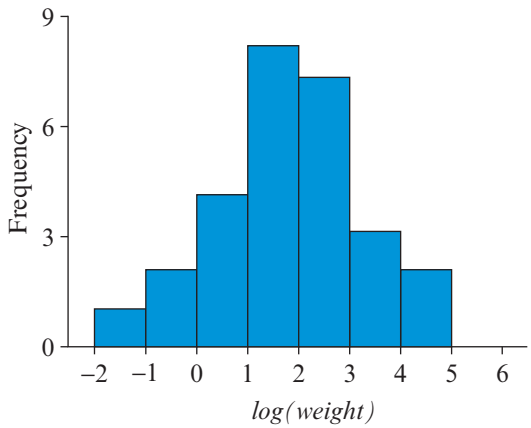
a Construct a histogram to display the distribution of brain weights and comment on its shape.

b Construct a histogram to display the log of the brain weights and note the shape of the distribution.

Interpreting a histogram with a log scale

Example 13

4 The histogram opposite shows the distribution of brain weights (in g) of 27 animal species plotted on a log scale.



a The brain weight (in g) of a mouse is 0.4 g. What value would be plotted on the log scale?

b The brain weight (in g) of an African elephant is 5712 g. What is the log of this brain weight (to two significant figures)?

c What brain weight (in g) is represented by the number 2 on the log scale?

d What brain weight (in g) is represented by the number –1 on the log scale?

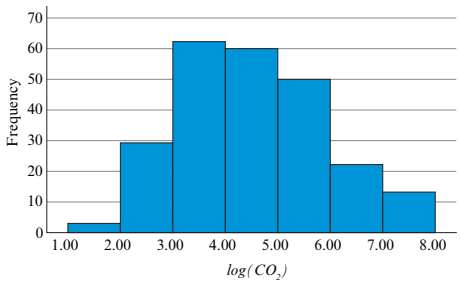
e Use the histogram to determine the number of these animals with brain weights:

i 1000 g or more **ii** from 1 to less than 100 g **iii** 1 g or more

ISBN 978-1-009-11041-9

Exam 1 style questions

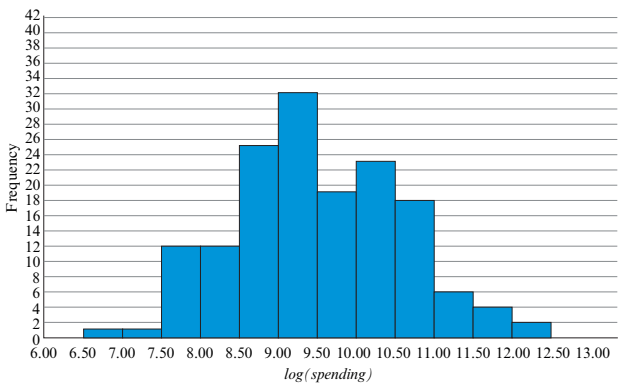
5 The histogram shows the carbon dioxide emissions (in thousands of metric tons) for 239 different countries, plotted on a $\log_{10}$ scale.



Based on this histogram, the percentage of countries with carbon dioxide emissions (in thousands of metric tons) from 10 000 to less than 100 000 is equal to:

A 21 **B** 25 **C** 26 **D** 50 **E** 60

6 The following histogram shows the amount spent by tourists from several countries in one year (*spending*), plotted on a $\log_{10}$ scale.



The number of countries where tourists spent from \$100 000 000 to less than \$1 000 000 000 per year is equal to:

A 12 **B** 25 **C** 33 **D** 37 **E** 51

1F Measures of centre and spread

Learning intentions

- To be able to understand the mean and the median as measures of centre.
- To be able to understand the range, interquartile range and standard deviation as measures of spread.
- To be able to know whether to use the median and interquartile range, or the mean and standard deviation, for a particular distribution.
- To be able to use a CAS calculator to calculate these summary statistics.

ISBN 978-1-009-11041-9

The median, range and interquartile range

The most versatile statistical tools for numerically describing the centre and spread of a distribution are:

- the **median** (the middle value) as a measure of **centre**;
- the **range** (the maximum spread of the data values), and the **interquartile range** (the spread of the middle half of data values) as measures of **spread**.

While these statistical values could be estimated only approximately from a histogram, they can be determined exactly when we use either a dot or stem plot.

Determining the median

We begin by revisiting the rule for locating the median of a data set.

The median

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

For n data values the median is located at the $\left(\frac{n+1}{2}\right)$th position.

When:

- n is odd, the median will be the middle data value
- n is even, the median will be the average of the two middle data values.

Example 14 Finding the median value in a data set

Order each of the following data sets, locate the median, and then write down its value.

a 2 9 1 8 3 5 3 8 1 **b** 10 1 3 4 8 6 10 1 2 9

Explanation

a For an odd number of data values, the median will be the middle data value.

1 Write down the data set in order.

2 Locate the middle data value by eye or use the rule.

3 Write down the median.

b For an even number of data values, the median will be the average of the two middle data values.

Solution

1 1 2 3 3 5 8 8 9

1 1 2 3 3 5 8 8 9

Median is the $\left(\frac{9+1}{2}\right)$th or fifth value.

Median = 3

ISBN 978-1-009-11041-9

Cambridge University Press

1 Write down the data set in order.	1 1 2 3 4 6 8 9 10 10 1 1 2 3 4 6 8 9 10 10
2 Locate the two middle data values and find their average or use the rule.	Median is the average of the 5th and 6th values.
Write down the median.	Median $= \left(\frac{4+6}{2}\right) = 5$

Note: You can check that you are correct by counting the number of data values each side of the median. They should be equal.

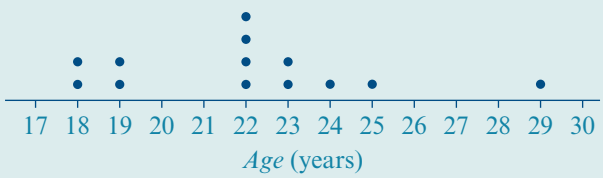
Using a dot plot to help locate medians

The process of calculating a median, as outlined above, is very simple in theory but can be time-consuming in practice. This is because you have to spend most of your time ordering the data set. For a very large data set this is a calculator task.

However, even for a reasonably large data set, locating a median in a dot or stem plot requires no more than counting because the data are already ordered for you.

Example 15 Finding the median value from a dot plot

The dot plot opposite displays the age distribution (in years) of the 13 members of a local cricket team.



Determine the median age of these cricketers and mark its location on the dot plot.

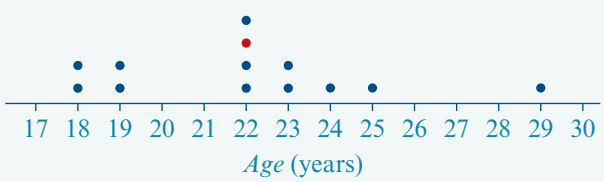
Explanation

The median value is the middle data value in the dot plot.

1 Locate the middle data value (or use the rule) and identify it on the dot plot.

2 Write down its value.

Solution



Median = 22 years

Example 16 Finding the median value from a stem plot

The stem plot opposite displays the maximum temperature (in °C) for 12 days in January.

Determine the median maximum temperature for these 12 days.

Key: 0|8 = 8°C

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

ISBN 978-1-009-11041-9

Explanation	Solution
1 For an even number of data values, as in this example, the median will be the average of the two middle data values. **2** Locate the two middle data values in the dot plot by eye (or use the rule) and identify them on the plot. **3** Determine the median by finding the average of these two data values.	Key: 0|8 = 8°C 1 | 8 9 9 2 | 0 2 5 7 8 9 9 3 | 1 3 $M = \frac{25 + 27}{2} = 26°C$

Having found the median value in a dot plot or stem plot, we now look at ways of doing the same with the first measure of spread, the range.

The range

The range

The range, R, is the simplest measure of spread of a distribution. It is the difference between the largest and smallest values in the data set.

R = largest data value − smallest data value

Example 17 Finding the range from a stem plot

The stem plot opposite displays the maximum temperature (in °C) for 12 days in January.

Determine the temperature range over these 12 days.

Key: 0|8 = 8°C

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

Explanation	Solution
1 Identify the lowest and highest values in the stem plot and write them down.	Key: 0|8 = 8°C 1 | 8 9 9 2 | 0 2 5 7 8 9 9 3 | 1 3
2 Substitute into the rule for the range and evaluate.	Lowest = 18, highest = 33, range = 33 − 18 = 15°C

Because the range depends only on the two extreme values in the data, it is not always an informative measure of spread. For example, one or other of these two values might be an outlier. Furthermore, any data with the same highest and lowest values will have the same range, irrespective of the way in which the data are spread out in between.

ISBN 978-1-009-11041-9

Cambridge University Press

A more refined measure of spread that overcomes these limitations of the range is the interquartile range (*IQR*).

The interquartile range (*IQR*)

Quartiles

To determine the value of the *IQR,* we first need to determine the quartiles.

Just as the median is the point that divides a distribution in half, **quartiles** are the points that divide a distribution into quarters. We use the symbols Q_1, Q_2 and Q_3 to represent the quartiles. Note that the second quartile, Q_2, is the median.

Determining the interquartile range

To find the interquartile range of a distribution:

- arrange all observations in order according to size
- divide the observations into two equal-sized groups, and if n is odd, omit the median from both groups
- locate Q_1, the **first quartile**, which is the median of the lower half of the observations
- locate Q_3, the **third quartile**, which is the median of the upper half of the observations.

The interquartile range *IQR* is then: $IQR = Q_3 - Q_1$

We can interpret the interquartile range as follows:

- Since Q_1, the first quartile, is the median of the lower half of the observations, then it follows that 25% of the data values are less than Q_1, and 75% are greater than Q_1.
- Since Q_3, the third quartile, is the median of the upper half of the observations, then it follows that 75% of the data values are less than Q_3, and 25% are greater than Q_3.
- Thus, the interquartile range (*IQR*) gives the spread of the middle 50% of data values.

Example 18 **Finding the *IQR* from an ordered stem plot when *n* is even**

Find the interquartile range of the weights of the 18 cats whose weights are displayed in the ordered stem plot below.

Weight (kg) 1 | 2 represents 1.2 kg

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

Explanation

1 There are 18 values in total. This means that there are nine values in the lower 'half', and nine in the upper 'half'.

Solution

ISBN 978-1-009-11041-9

Cambridge University Press

2 The median of the lower half (Q_1) is the middle of lower nine values, which is the 5th value from the bottom.	Lower half: 1.9 2.1 2.3 2.5 2.8 3.0 3.0 3.4 3.9 $Q_1 = 2.8$
3 The median of the upper half (Q_3) is the middle of the upper nine values, which is the 5th value from the top.	Upper half: 3.9 4.0 4.4 4.5 4.8 5.0 5.3 6.3 6.4 $Q_3 = 4.8$
4 Determine the IQR using $IQR = Q_3 - Q_1$.	$IQR = Q_3 - Q_1 = 4.8 - 2.8 = 2.0$

Example 19 Finding the *IQR* from an ordered stem plot when *n* is odd

The stem plot shows the life expectancy (in years) for 23 countries. Find the IQR for life expectancies.

Stem: 5|2 = 52 years

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

Explanation

1 Since there are 23 values, the median is the 12th value from either end which is 73. Mark the value 73 on the stem plot.

2 Since n is odd, to find the quartiles the median value is excluded. This leaves 11 values below the median and 11 values above the median. Then:

- Q_1 = midpoint of the bottom 11 data values
- Q_3 = midpoint of the top 11 data values.

Write these values down.

3 Determine the IQR using $IQR = Q_3 - Q_1$.

Solution

Stem: 5|2 = 52 years

5	2
5	5 6
6	4 (Q_1)
6	6 6 7 9 (median)
7	1 2 2 3 3 4 4 4 4
7	5 5 6 6 7 7 (Q_3)

$Q_1 = 66$, $Q_3 = 75$

$IQR = Q_3 - Q_1 = 75 - 66 = 9$

To check that these quartiles are correct, write the data values in order, and mark the median and the quartiles. If correct, the median divides the data set up into four equal groups.

ISBN 978-1-009-11041-9

Cambridge University Press

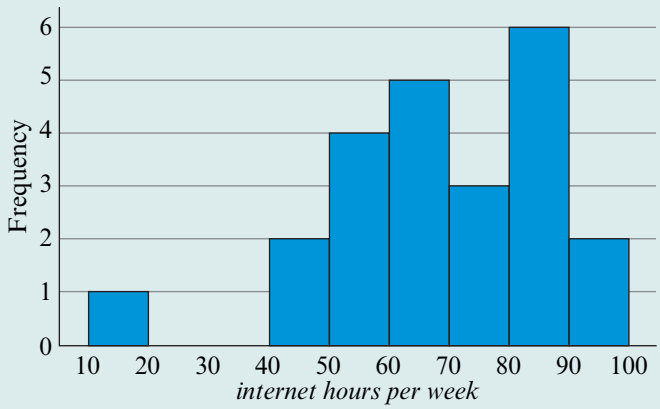
Q_1 Q_2 $(= M)$ Q_3

52 55 56 64 66 **66** 67 69 71 72 72 **73** 73 74 74 74 74 **75** 75 76 76 77 77

5 values 5 values 5 values 5 values

Example 20 Finding the median and quartiles from a histogram

The histogram shows the average number of hours per week a group of 23 people spent on the internet. Find possible values for the median and quartiles of this distribution.



Explanation

1 Since there are 23 values, we can locate which interval contains the median by adding the number of values in each interval moving from left to right.

- M = the 12th value from the bottom (adding from left to right)

2 Similarly

- Q_1 = the 6th value from the bottom (adding from left to right)
- Q_3 = the 6th value from the top (adding from right to left)

Solution

There is 1 value in the interval 10-20 (total 1), 2 values in the interval 40-50 (total 3), 4 values in the interval 50-60 (total 7), 5 values in the interval 60-70 (total 12). Thus the median is in the interval 60-70.

Q_1 is in the interval 50-60

Q_3 is in the interval 80-90

Why is the *IQR* a more useful measure of spread than the range?

The *IQR* is a measure of spread of a distribution that includes the middle 50% of observations. Since the upper 25% and lower 25% of observations are discarded, the interquartile range is generally not affected by the presence of outliers.

ISBN 978-1-009-11041-9

The mean and standard deviation

So far, we have looked at methods for describing the centre and spread for distributions of any shape. We used the median, *IQR* and range for this purpose. In this section, we will look at alternative measures of centre (the mean) and spread (the standard deviation) that are only useful when working with symmetric distributions without outliers. While this may seem unnecessarily restrictive, these two measures have the advantage of being able to fully describe the centre and spread of a symmetric distribution with only two numbers.

The mean

The **mean** of a set of data is what most people call the 'average'. The mean of a set of data is given by:

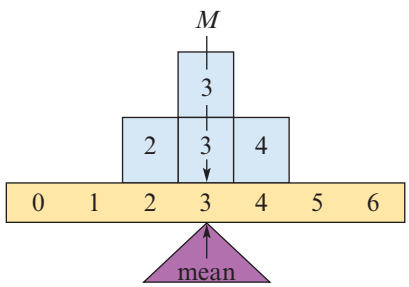
$$\text{mean} = \frac{\text{sum of data values}}{\text{total number of data values}}$$

For example, consider the set of data:

2 3 3 4

The mean of this set of data is given by:

$$\text{mean} = \frac{2+3+3+4}{4} = \frac{12}{4} = 3$$



From a pictorial point of view, the mean is the balance point of a distribution (see above).

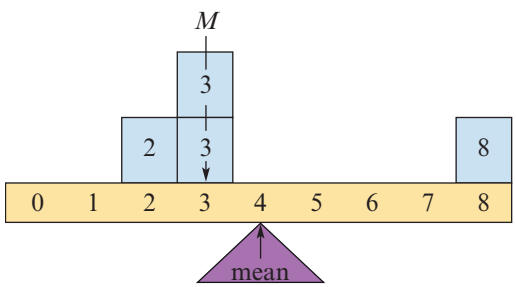
Note that in this case, the mean and the median coincide; the balance point of the distribution is also the point that splits the distribution in half. That is, there are two data points to the left of the mean and two to the right. This is a general characteristic of symmetric distributions.

However, consider the data set

2 3 3 8

The median remains at $M = 3$, but:

$$\text{mean} = \frac{2+3+3+8}{4} = \frac{16}{4} = 4$$



Note that the mean is affected by changing the largest data value but that the median is not.

Some notation

Because the rule for the mean is relatively simple, it is easy to write in words. However, later you will meet other rules for calculating statistical quantities that are extremely complicated and hard to write out in words.

To overcome this problem, we introduce a shorthand notation that enables complex statistical formulas to be written out in a compact form. In this notation, we use:

ISBN 978-1-009-11041-9

Cambridge University Press

- the Greek capital letter sigma, $\sum$, as a shorthand way of writing 'sum of'
- a lower case x to represent a data value
- a lower case x with a bar, $\bar{x}$ (pronounced 'x *bar*'), to represent the mean of the data values
- an n to represent the total number of data values.

The rule for calculating the mean then becomes: $\bar{x} = \dfrac{\sum x}{n}$

Example 21 **Calculating the mean from the formula**

The following is a set of reaction times (in milliseconds): 38 36 35 43 46 64 48 25

Write down the values of the following, correct to one decimal place.

a n **b** $\sum x$ **c** $\bar{x}$

Explanation

a n is the number of data values.

b $\sum x$ is the sum of the data values.

c $\bar{x}$ is the mean. It is defined by $\bar{x} = \dfrac{\sum x}{n}$.

Solution

$n = 8$

$\sum x = 38 + 36 + 35 + 43 + 46 + 64 + 48 + 25$

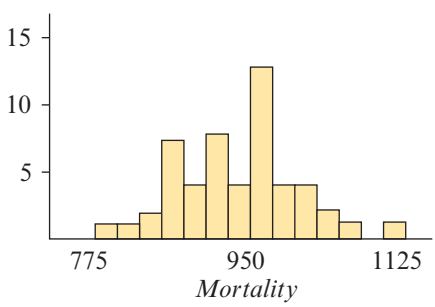
$= 335$

$\bar{x} = \dfrac{\sum x}{n} = \dfrac{335}{8} = 41.9$

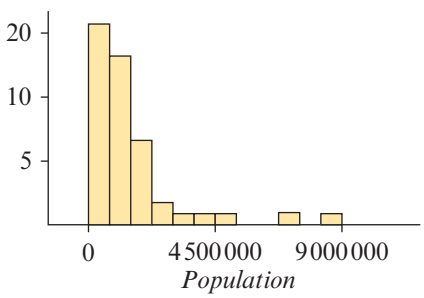
The relationship between the mean and the median

Whereas the median lies at the midpoint of a distribution, the mean is the balance point of the distribution. For approximately symmetric distributions, both the median and mean will be approximately equal in value.

An example of a symmetric distribution is the distribution of mortality rates for 60 US cities shown opposite. Calculations reveal that the mean mortality rate for the cities is 940 per 100 000 while the median mortality rate is 944 per 100 000 people. As expected, the mean and median are approximately equal in value.



An example of a highly skewed distribution is the population distribution of different cities, shown opposite. This distribution is clearly positively skewed with two outliers. The mean population is 1.4 million, while the median population is 0.9 million. They are quite different in value. The mean has been affected by the extreme values in the tail and no longer represents the typical city.



ISBN 978-1-009-11041-9

When to use the median rather than the mean

Because the value of the median is relatively unaffected by the presence of extreme values in a distribution, it is said to be a **resistant** statistic. For this reason, the median is frequently used as a measure of centre when the distribution is known to be clearly skewed and/or likely to contain outliers.

For example, median house prices are used to compare housing prices between capital cities in Australia because the distribution of house prices tends to be positively skewed. There are always a small number of very expensive houses sold for much higher prices than the rest of houses sold.

However, if a distribution is symmetric, there will be little difference in the value of the mean and median and we can use either. In such circumstances, the mean is often preferred because:

- it is more familiar to most people
- more can be done with it theoretically, particularly in the area of statistical inference (which you will learn about if you are doing Mathematics Methods).

Choosing between the mean and the median

The mean and the median are both measures of the centre of a distribution. If the distribution is:

- symmetric and there are no outliers, either the mean or the median can be used to indicate the centre of the distribution
- clearly skewed and/or there are outliers, it is more appropriate to use the median to indicate the centre of the distribution.

The standard deviation

To measure the spread of a data distribution around the median (M) we use the interquartile range (IQR). To measure the spread of a data distribution about the mean ($\bar{x}$) we use the **standard deviation** (s).

The standard deviation

The formula for the standard deviation, s, is: $s = \sqrt{\dfrac{\sum(x - \bar{x})^2}{n - 1}}$

Although not easy to see from the formula, the standard deviation is an average of the squared deviations of each data value from the mean. We work with the squared deviations because the sum of the deviations around the mean (the balance point) will always be zero.

Calculating the standard deviation

Normally, you will use your calculator to determine the value of a standard deviation. Instructions for the TI-Nspire or ClassPad follow.

ISBN 978-1-009-11041-9

Cambridge University Press

CAS 3: How to calculate the mean and standard deviation using the TI-Nspire CAS

The following are the heights (in cm) of a group of women.

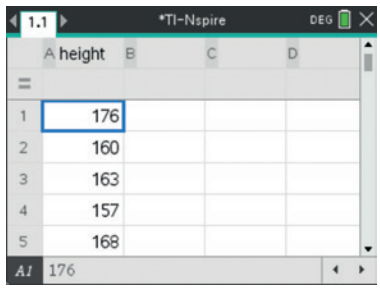
176 160 163 157 168 172 173 169

Determine the mean and standard deviation of the women's heights. Give your answers correct to two decimal places.

Steps

1 Start a new document by pressing ctrl + N.

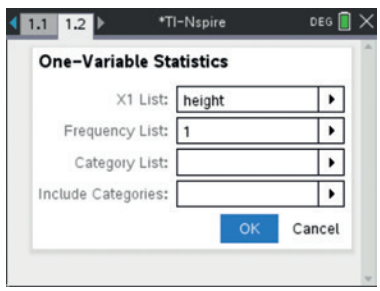
2 Select **Add Lists & Spreadsheet**.
Enter the data into a list named *height*, as shown.



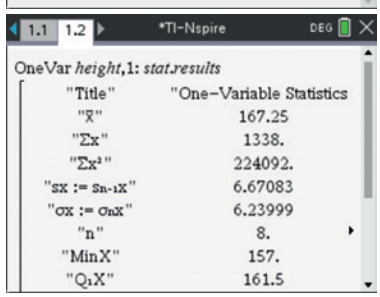
3 Statistical calculations can be done in either the **Lists & Spreadsheet** application or the **Calculator** application (used here).
Press ctrl + I and select **Add Calculator**.

a Press menu>**Statistics>Stat Calculations>One-Variable Statistics**. Press enter to accept the **Num of Lists** as 1.

b **i** To complete this screen, use the ▶ arrow and enter to paste in the list name *height*.



ii Pressing enter exits this screen and generates the results screen shown opposite.



4 Write down the answers to the required degree of accuracy (i.e. two decimal places).

The mean height of the women is $\bar{x} = 167.25$ cm and the standard deviation is $s = 6.67$ cm.

Notes: **a** The sample standard deviation is **sx**.
b Use the ▲ ▼ arrows to scroll through the results screen to obtain values for additional statistical values.

ISBN 978-1-009-11041-9

Cambridge University Press

CAS 3: How to calculate the mean and standard deviation using the ClassPad

The following are all heights (in cm) of a group of women.

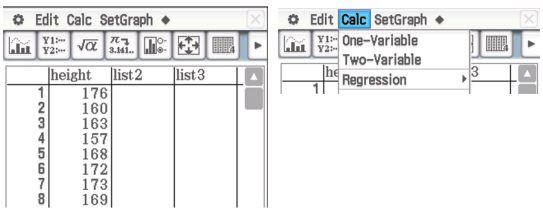
176 160 163 157 168 172 173 169

Determine the mean and standard deviation of the women's heights correct to two decimal places.

Steps

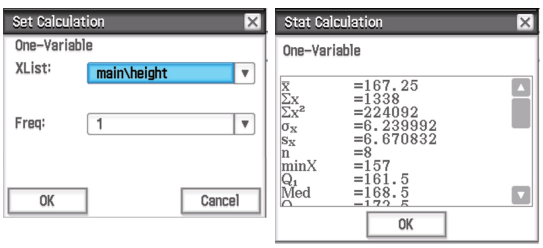
1 Open the **Statistics** application 

and enter the data into the column labelled *height*.



2 To calculate the mean and standard deviation, select **Calc** from the menu One-Variable from the drop-down menu to open the **Set Calculation** dialog box shown below.

3 Complete the dialog box as shown.
- **XList**: select **main\height** (▼).
- **Freq**: leave as **1**.



4 Tap **OK** to confirm your selections and calculate the required statistics, as shown.

5 Write down the answers to two decimal places.

The mean height of the women is $\bar{x} = 167.25$ cm.

The standard deviation is $s_x = 6.67$ cm.

Notes: **a** The value of the standard deviation is given by s_x.

b Use the side-bar arrows to scroll through the results screen to obtain values for additional statistical values (i.e. median, Q_3 and the maximum value) if required.

Exercise 1F

Determining the median from data

Example 14

1 Locate the medians of the following data sets. For each set of data, check that the median divides the ordered data set into two equal groups.

a 4 9 3 1 8 6

b 10 9 12 20 14

ISBN 978-1-009-11041-9

Cambridge University Press

2 The prices of nine second-hand mountain bikes advertised for sale were as follows.

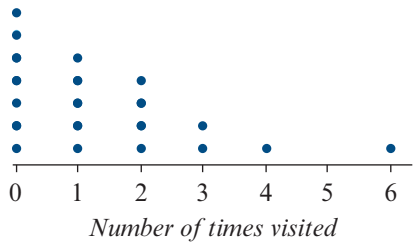
$650 $3500 $750 $500 $1790 $1200 $2950 $430 $850

What is the median price of these bikes?

Determining the median from a dot plot

Example 15

3 The dot plot opposite displays the number of times 20 shoppers visited their supermarket in a week. Find the median number of visits.



Determining the median and range from a stem plot

4 The following stem plot shows the distribution of the time it took (in minutes) for each of a group of 25 people to solve a complex task.

Time (minutes) key: 4|0 represents 4.0

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

Example 16

a Find the median time taken.

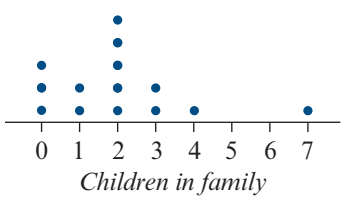
Example 17

b Find the range of the time taken.

Determining the median and quartiles from a dot plot

Example 18

5 The dot plot shows the distribution of the number of children in each of 14 families.



a Determine the median, M.

b Determine the quartiles Q_1 and Q_3.

c Calculate the IQR.

d Calculate the range, R.

e By writing the data values in a line, check that the quartiles and the median have divided the data set up into four equal groups.

ISBN 978-1-009-11041-9

Cambridge University Press

Determining the median and quartiles from a stem plot

6 The stem plot displays the infant mortality rates (deaths per 1000 live births) in 14 countries.

Key: 0|7 = 7

0	7 7 9
1	0 0 0 0 2 2 4
1	5
2	0 1
2	5

a Determine the median, M.

b Determine the quartiles Q_1 and Q_3.

c Calculate the IQR and the range, R.

7 The stem plot displays the test scores for 20 students.

Key: 1|0 = 10

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

a Describe the shape of the distribution.

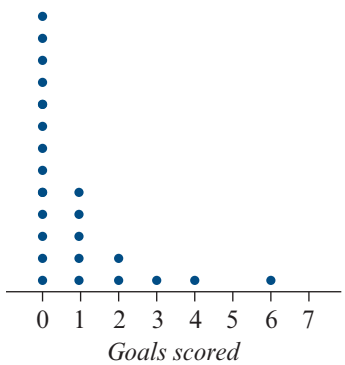
b Determine the median, M.

c Determine the quartiles, Q_1 and Q_3.

d Calculate the IQR and the range, R.

Example 19

8 The dot plot displays the number of goals scored in 23 games.



a Describe the shape of the distribution and note outliers (if any).

b Without using your calculator determine:

i the median, M.

ii the IQR.

9 The stem plot displays the university participation rates (%) in 17 countries.

Key: 0|1 = 1

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

a Determine the median, M.

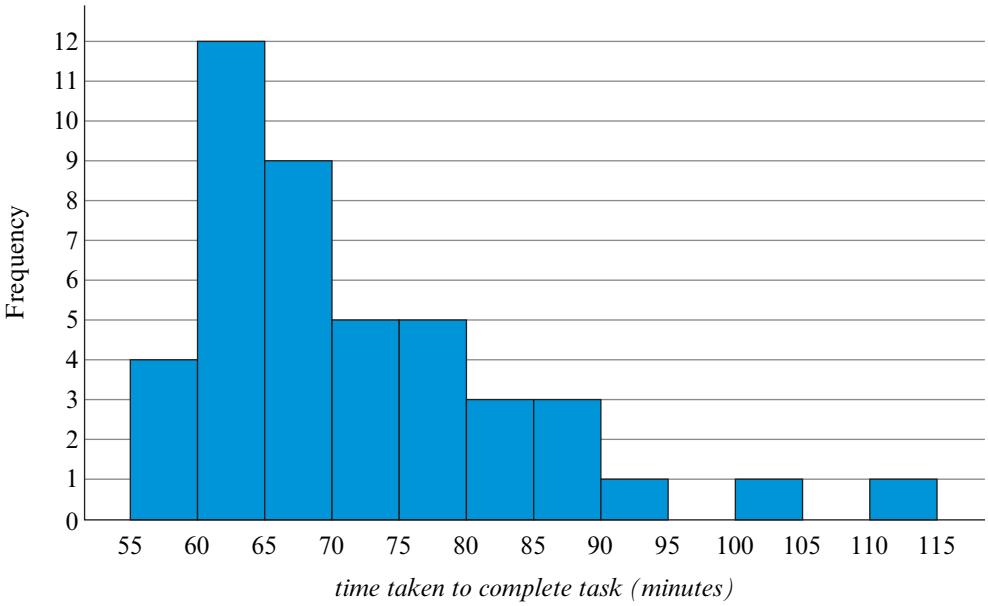
b Determine the quartiles Q_1 and Q_3.

c Calculate the IQR and the range, R.

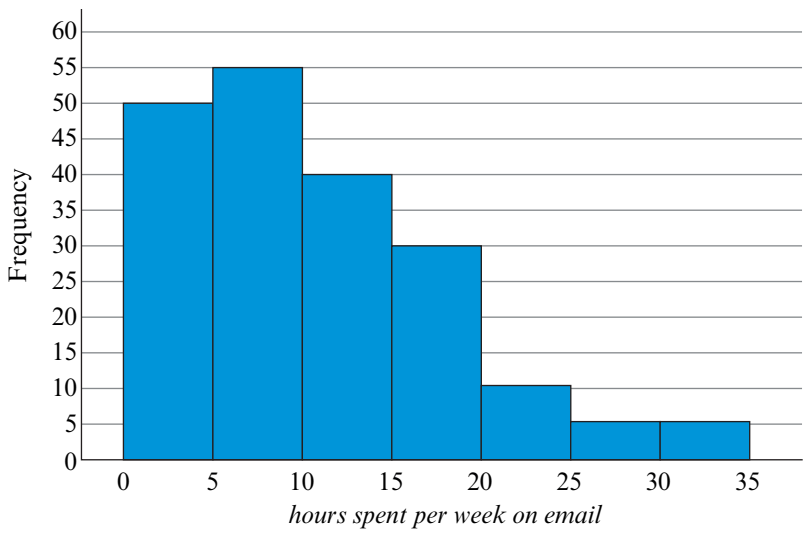
ISBN 978-1-009-11041-9

Cambridge University Press

Example 20 **10** The histogram shows the time taken to complete a complex task by a group of students. Find possible values for the median and quartiles of this distribution.



11 A group of 195 people were asked to record (to one decimal place) the average number of hours they spent on email each week over a 10 week period. The data are shown in the following histogram:



a Find possible values for the median.

b Find the maximum value for the *IQR*.

ISBN 978-1-009-11041-9

Determining the mean, median and mode from data

Example 21

12 For each of the following data sets, write down the value of n, the value of Σx and hence evaluate $\bar{x}$.

a 2 5 2 3 **b** 12 15 20 32 25 **c** 2 1 3 2 5 3 5

13 Calculate the mean and locate the median and modal value(s) of the following scores.

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

b 3 12 5 4 3 2 6 5 4 5 5 6

14 The temperature of a hospital patient (in degrees Celsius) taken at 6-hourly intervals over 2 days was as follows.

35.6 36.5 37.2 35.5 36.0 36.5 35.5 36.0

a Calculate the patient's mean and median temperature over the 2-day period.

b What do these values tell you about the distribution of the patient's temperature?

15 The amounts (in dollars) spent by seven customers at a corner store were:

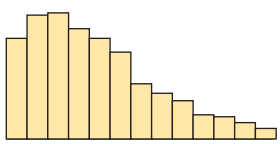
0.90 0.80 2.15 16.55 1.70 0.80 2.65

a Calculate the mean and median amount spent by the customers.

b Does the mean or the median give the best indication of the typical amount spent by customers? Explain your answer.

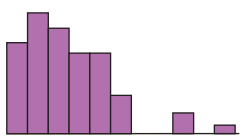
16 For which of the following distributions might you question using the mean as a measure of the centre of the distribution? Justify your selection.

a



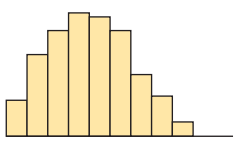
Age distribution in a country

b



Urban car accident rates

c



Blood cholesterol levels

17 The stem plot shows the distribution of weights (in kg) of 22 footballers.

a Name the shape of the distribution. Which measure of centre, the mean or the median, do you think would best indicate the typical weight of these footballers?

b Determine both the mean and median to check your prediction.

Weight (kg)

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

ISBN 978-1-009-11041-9

Cambridge University Press

The concept of standard deviation

18 Which measure of spread:

a always incorporates 50% of the scores?

b uses only the smallest and largest scores in the distribution?

c gives the average variation around the mean?

19 Without using the statistical capabilities of your calculator, write down the mean and standard deviation of the following six data values: 7.1 7.1 7.1 7.1 7.1 7.1

20 For which of the following variables does it *not* make sense to calculate a mean or standard deviation?

a Speed (in km/h) **b** Sex **c** Age (in years)

d Post code **e** Neck circumference (in cm)

f Weight (underweight, normal, overweight)

Calculating the mean and standard deviation using a CAS calculator

21 A sample of 10 students were given a general knowledge test with the following results.

20 20 19 21 21 18 20 22 23 17

a Calculate the mean and standard deviation of the test scores, correct to one decimal place.

b The median test score is 20, which is similar in value to the mean. What does this tell you about the distribution of test scores?

22 Calculate the mean and standard deviation for the variables in the table.
Give answers to the nearest whole number for cars and TVs, and one decimal place for alcohol consumption.

Number of TVs/ 1000	*Number of cars/ 1000*	*Alcohol consumption (litres)*
378	417	17.6
404	286	12.5
471	435	16.0
354	370	24.1
539	217	9.9
381	357	9.5
624	550	14.6

ISBN 978-1-009-11041-9

Exam 1 style questions

Use the following information to answer questions 23 and 24

The stem plot displays the number of times each person in a sample of 16 people bought a take-away coffee in the last month.

Key: 0|1 = 1

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

23 The median, M, of the number of take away coffees bought is equal to:

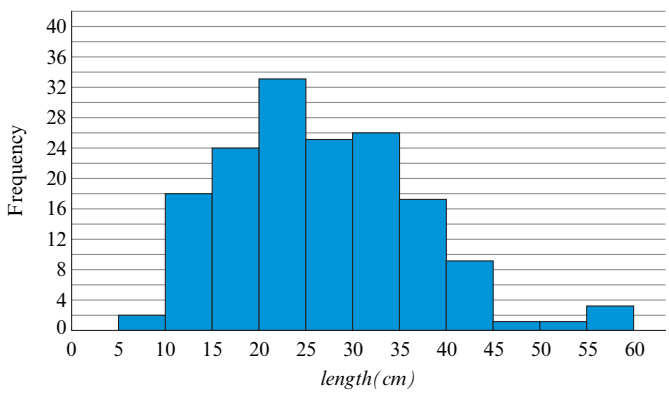
A 10.5 **B** 18.5 **C** 20.5 **D** 21 **E** 26

24 The interquartile range, IQR, of the number of take away coffees bought is equal to:

A 10.5 **B** 15.0 **C** 15.5 **D** 18.5 **E** 20.5

Use the following information to answer questions 25 and 26

The following histogram shows the length (in cm) for each of 159 fish.



25 The median fish length (in cm) could be

A 15.8 **B** 24.3 **C** 25.2 **D** 31.4 **E** 80.0

26 The first quartile (Q_1) for this distribution could be

A 12.2 **B** 16.7 **C** 29.0 **D** 20.2 **E** 25.0

ISBN 978-1-009-11041-9

1G The five-number summary and the boxplot

Learning intentions

- To be able to construct the boxplot for displaying the distribution of a numerical data.
- To be able to define and identify **outliers**.
- To be able to construct both **simple boxplots** and **boxplots with outliers**.
- To be able to determine the features of a distribution from a boxplot.
- To be able to use a CAS calculator to construct a boxplot.

The five-number summary

Knowing the median and quartiles tells us quite a lot about the centre and spread of the distribution. If we also knew something about the tails (ends) we would have a good picture of the whole distribution. This can be achieved by recording the smallest and largest values of the data set. Putting all this information together gives the **five-number summary**.

Five-number summary

A listing of the median, M, the quartiles Q_1 and Q_3, and the smallest and largest data values of a distribution, written in the order

minimum, Q_1, M, Q_3, maximum

is known as a five-number summary.

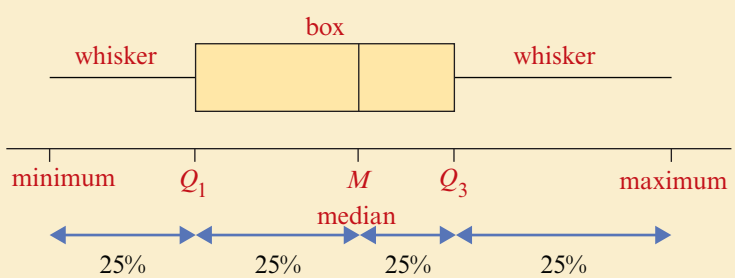
The five-number summary is the starting point for constructing one of the most useful graphical tools in data analysis, the boxplot.

The boxplot

The **boxplot** (or box-and-whisker plot) is a graphical display of a five-number summary. The essential features of a boxplot are summarised below.

The boxplot

A boxplot is a graphical display of a five-number summary.



ISBN 978-1-009-11041-9

In a boxplot:

- a box extends from Q_1 to Q_3, locating the middle 50% of the data values
- the median is shown by a vertical line drawn within the box
- lines (called whiskers) are extended out from the lower and upper ends of the box to the smallest and largest data values of the data set respectively
- 25% of the data values are from the minimum to Q_1
- 25% of the data values are from Q_1 to the median M
- 25% of the data values are from the median M to Q_3
- 25% of the data values are from Q_3 to the maximum

Example 22 Constructing a boxplot from a five-number summary

The stem plot shows the distribution of life expectancies (in years) in 23 countries.

Key: 5|2 = 52 years

5	2	← minimum
5	5 6	
6	4	
6	6 6 7 9	← Q_1, median
7	1 2 2 3 3 4 4 4 4	
7	5 5 6 6 7 7	← Q_3, maximum

The five-number summary for these data is:

minimum	52
first quartile (Q_1)	66
median (M)	73
third quartile (Q_3)	75
maximum	77

Use the five-number summary to construct a boxplot.

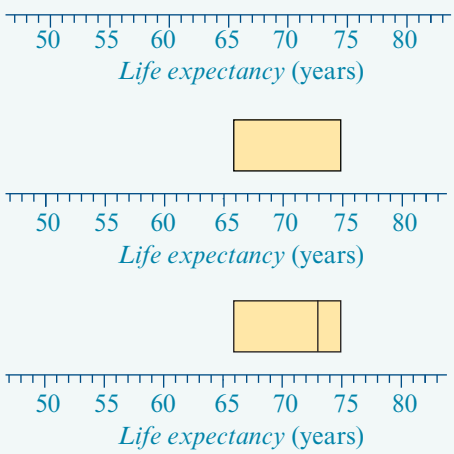
Explanation

1 Draw a labelled and scaled number line that covers the full range of values.

2 Draw a box starting at $Q_1 = 66$ and ending at $Q_3 = 75$.

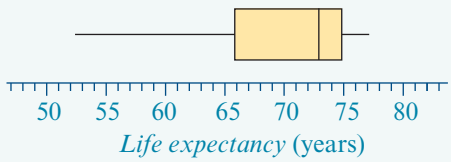
3 Mark the median value with a vertical line segment at $M = 73$.

Solution



ISBN 978-1-009-11041-9

4 Draw the whiskers: lines joining the midpoint of the ends of the box to the minimum and maximum values, 52 and 77.



Boxplots with outliers

An extension of the boxplot can also be used to identify possible outliers in a data set.

Sometimes it is difficult to decide whether or not an observation is an outlier. For example, a boxplot might have one extremely long whisker. How might we explain this?

- One explanation is that the data distribution is extremely skewed with lots of data values in its tail.
- Another explanation is that the long whisker hides one or more outliers.

By modifying the boxplots, we can decide which explanation is most likely, but firstly we need a more exact definition of an outlier.

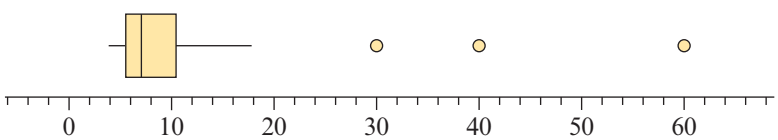
Defining outliers

Outlier

An **outlier** in a distribution is any data point that lies more than 1.5 interquartile ranges below the first quartile or more than 1.5 interquartile ranges above the third quartile.

To be more informative the boxplot can be modified so that the outliers are plotted individually in the boxplot with a dot or cross, and the whisker now ends only to the largest or smallest data value that is not outside these limits.

An example of a boxplot with outliers is shown opposite.



Three of the data values 30, 40, and 60 are possible outliers.

To display outliers on a boxplot, we must first determine the location of what we call the **upper** and **lower fences**. These are imaginary lines drawn one and a half interquartile ranges (or box widths) above and below the box ends, as shown in the diagram following. Data values outside these fences are then classified as possible outliers and plotted separately.

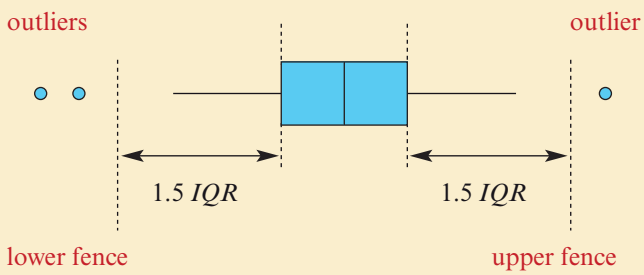
ISBN 978-1-009-11041-9

Cambridge University Press

Using a boxplot to display outliers

In a boxplot, possible outliers are defined as being those values that are:

- greater than $Q_3 + 1.5 \times IQR$ (upper fence)
- less than $Q_1 - 1.5 \times IQR$ (lower fence).



When drawing a boxplot, any observation identified as an outlier is shown by a dot. The whiskers end at the smallest and largest values that are not classified as outliers.

While we have used a five-number summary as the starting point for our introduction to boxplots, in practice the starting point for constructing a boxplot is raw data. Constructing a boxplot from raw data is a task for your CAS calculator.

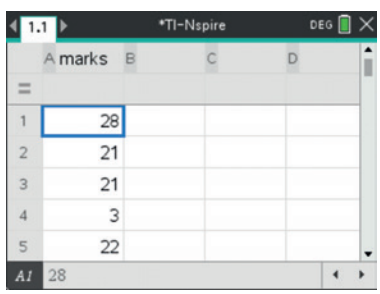
CAS 4: How to construct a boxplot with outliers using the TI-Nspire CAS

Display the following set of 19 marks in the form of a boxplot with outliers.

28 21 21 3 22 31 35 26 27 33
43 31 30 34 48 36 35 23 24

Steps

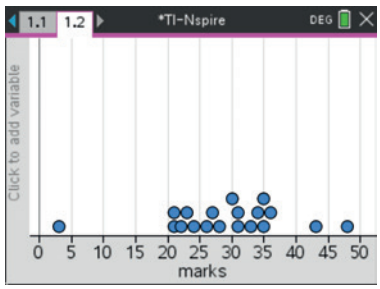
1. Start a new document by pressing ctrl+N.
2. Select **Add Lists & Spreadsheet**. Enter the data into a list called ***marks*** as shown.
3. Statistical graphing is done through the **Data & Statistics** application.
   Press ctrl + I and select **Add Data & Statistics**.



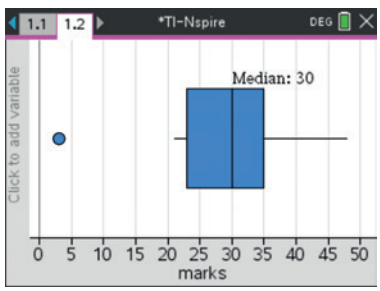
ISBN 978-1-009-11041-9

Note: A random display of dots will appear – this indicates that list data are available for plotting. Such a dot is not a statistical plot.

a Click on the **Click to add variable** on the x-axis and select the variable ***marks***. A dot plot is displayed by default as shown opposite.



b To change the plot to a boxplot press menu>**Plot Type**>**boxplot**. Your screen should now look like that shown opposite.



4 Data analysis

Key values can be read from the boxplot by moving the cursor over the plot or using menu>**Analyze>Graph Trace**.

Starting at the far left of the plot, we see that the:

- minimum value is 3 (an outlier)
- first quartile is 23 ($\mathbf{Q_1 = 23}$)
- median is 30 (**Median = 30**)
- third quartile is 35 ($\mathbf{Q_3 = 35}$)
- maximum value is 48.

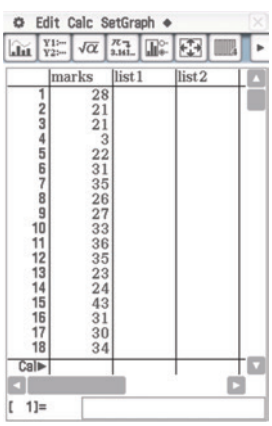
CAS 4: How to construct a boxplot with outliers using the ClassPad

Display the following set of 19 marks in the form of a boxplot with outliers.

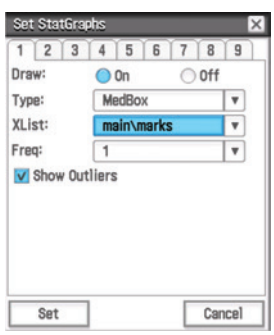
28 21 21 3 22 31 35 26 27 33
43 31 30 34 48 36 35 23 24

Steps

1 Open the **Statistics** application Statistics and enter the data into the column labelled **marks**.



2 Open the **Set StatGraphs** dialog box by tapping in the toolbar. Complete the dialog box as shown below.



- **Draw**: select **On**.
- **Type**: select **MedBox** (▼).
- **XList**: select **main\marks** (▼).
- **Freq**: leave as **1**.

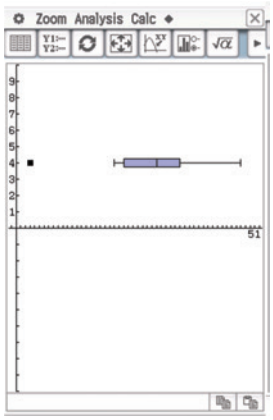
Tap the **Show Outliers** box to tick (☑).

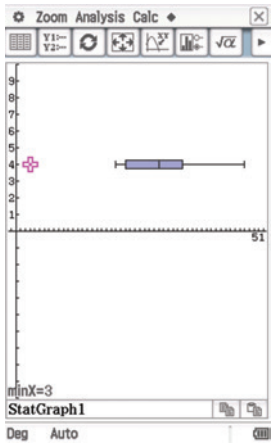
ISBN 978-1-009-11041-9

3 Tap **Set** to confirm your selections and plot the boxplot b tapping . The graph is drawn in an automatically scaled window, as shown.

4 Tap the Resize icon at the bottom of the screen for a full-screen graph.

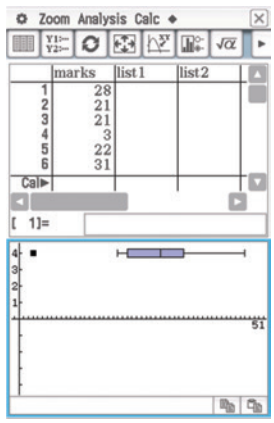
Note: If you have more than one graph on your screen, tap the data screen, select **StatGraph** and turn off any unwanted graphs.





5 Tap to read key values. This places a marker on the boxplot (+), as shown. Use the horizontal cursor arrows (◀) and (▶) to move from point to point on the boxplot. We see that the:

- minimum value is 3 (**minX = 3**; an outlier)
- first quartile is 23 ($\mathbf{Q_1 = 23}$)
- median is 30 (**Med = 30**)
- third quartile is 35 ($\mathbf{Q_3 = 35}$)
- maximum value is 48 (**maxX = 48**).

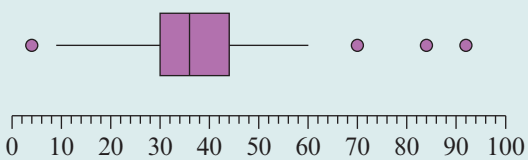


Interpreting boxplots

Constructing a boxplot is not an end in itself. The prime reason to construct boxplots is to help us answer statistical questions. To do this, you need to know how to read values from a boxplot and use them to determine statistics such as the median, the interquartile range and the range. We also use boxplots to identify possible outliers.

Example 23 Reading values from a boxplot

For the boxplot shown, write down the values of:



a the median

b the quartiles Q_1 and Q_3

c the interquartile range (IQR)

d the minimum and maximum values

e the values of any possible outliers

f the smallest value in the upper end of the data set that will be classified as an outlier

g the largest value in the lower end of the data set that will be classified as an outlier.

ISBN 978-1-009-11041-9

Cambridge University Press

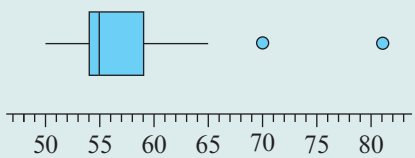
Explanation	Solution
a The median (the vertical line in the box)	$M = 36$
b Quartiles Q_1 and Q_3 (end points of box)	$Q_1 = 30$, $Q_3 = 44$
c Interquartile range ($IQR = Q_3 - Q_1$)	$IQR = Q_3 - Q_1 = 44 - 30 = 14$
d Minimum and maximum values (extremes)	$Min = 4$, $Max = 92$
e The values of the possible outliers (dots)	4, 70, 84 and 92 are possible outliers
f Upper fence (given by $Q_3 + 1.5 \times IQR$)	Upper fence $= Q_3 + 1.5 \times IQR$ $= 44 + 1.5 \times 14 = 65$ Any value above 65 is an outlier.
g Lower fence (given by $Q_1 - 1.5 \times IQR$)	Lower fence $= Q_1 - 1.5 \times IQR$ $= 30 - 1.5 \times 14 = 9$ Any value below 9 is an outlier.

Once we know the location of the quartiles, we can use the boxplot to estimate percentages.

Example 24 Estimating percentages from a boxplot

For the boxplot shown, estimate the percentage of values:

a less than 54 **b** less than 55
c less than 59 **d** greater than 59
e between 54 and 59 **f** between 54 and 86.



Explanation	Solution
a 54 is the first quartile (Q_1); 25% of values are less than Q_1.	**a** 25%
b 55 is the median or second quartile (Q_2); 50% of values are less than Q_2.	**b** 50%
c 59 is the third quartile (Q_3); 75% of values are less than Q_3.	**c** 75%
d 75% of values are less than 59 and 25% are greater than 59.	**d** 25%
e As 75% of values are less than 59 and 25% are less than 54, 50% of values are between 54 and 59.	**e** 50%
f As 100% of values are less than 86 and 25% of values are less than 54, 75% of values are between 54 and 86.	**f** 75%

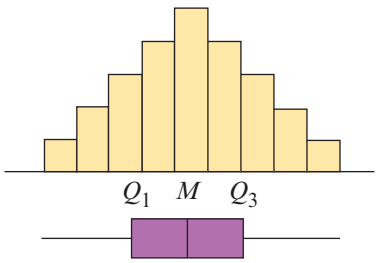
ISBN 978-1-009-11041-9

Relating a boxplot to shape

When there are a reasonable number of data values, the shape of a distribution can be identified from a boxplot.

A symmetric distribution

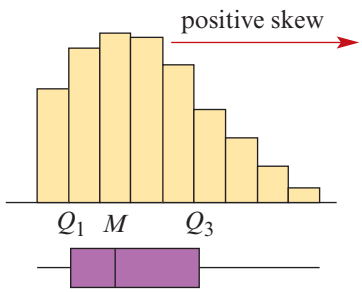
A symmetric distribution tends to be centred on its median and have values evenly spread around the median. As a result, its boxplot will also be symmetric, its median is close to the middle of the box and its whiskers are approximately equal in length.



Positively skewed distributions

Positively skewed distributions are characterised by a cluster of data values around the median at the left-hand end of the distribution with a gradual tailing off to the right.

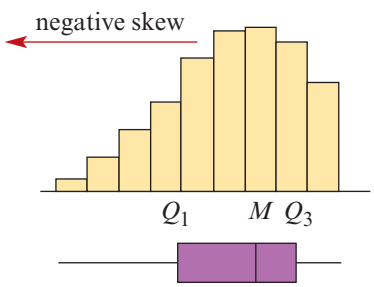
As a result, the boxplot of a positively skewed distribution will have its median off-centre and to the left-hand side of its box. The left-hand whisker will be short, while the right-hand whisker will be long, reflecting the gradual tailing off of data values to the right.



Negatively skewed distributions

Negatively skewed distributions are characterised by a clustering of data values around the median at the right-hand end of the distribution, with a gradual tailing off of data values to the left.

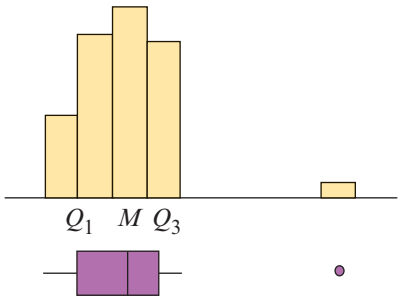
As a result, the boxplot of a negatively skewed distribution has the median off-centre and in the right-hand side of its box. The right-hand whisker will be short, while the left-hand whisker will be long, reflecting the gradual tailing off of data values to the left.



ISBN 978-1-009-11041-9

Distributions with outliers

Distributions with outliers are characterised by large gaps between the main body and data values in the tails. The histogram opposite displays a distribution with an outlier. In the corresponding boxplot, the box and whiskers represent the main body of data and the dot, separated by a gap from the box and whiskers, an outlier.

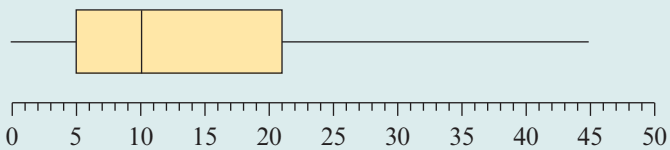


Using boxplots to describe a distribution

Because of the wealth of information contained in a boxplot, it is an extremely powerful tool for describing the features of distribution in terms of shape, centre, spread and outliers.

Example 25 **Using a boxplot to describe the features of distribution without outliers**

Describe the distribution represented by the boxplot in terms of shape, centre and spread. Give appropriate values.

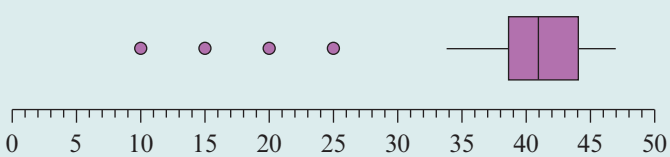


Solution

The distribution is positively skewed with no outliers. The distribution is centred at 10, the median value. The spread of the distribution, as measured by the *IQR*, is 16 and, as measured by the range, 45.

Example 26 **Using a boxplot to describe the features of a distribution with outliers**

Describe the distributions represented by the boxplot in terms of shape and outliers, centre and spread. Give appropriate values.



Solution

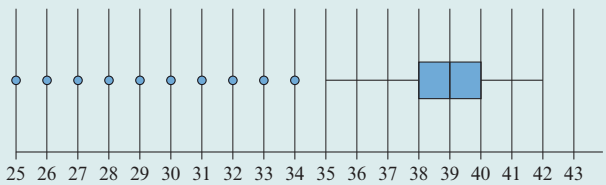
The distribution is symmetric but with outliers. The distribution is centred at 41, the median value. The spread of the distribution, as measured by the *IQR*, is 5.5 and, as measured by the range, 37. There are four outliers: 10, 15, 20 and 25.

ISBN 978-1-009-11041-9

Earlier in this chapter we attempted to describe the feature of a distribution from a histogram. We found that from the histogram it was difficult to give exact values for centre and spread, and to clearly identify outliers. This is much easier to do from a boxplot.

Example 27 Using a boxplot to answer statistical questions

The boxplot shows the gestation period (completed weeks) for a sample for 1000 babies born in Australia one year. Describe the distribution of gestation period in terms of shape, centre, spread and outliers.



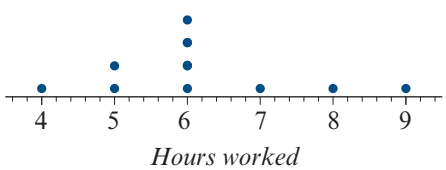
Solution

The distribution of gestational period is negatively skewed with several outliers. The distribution is centred at 39 weeks, the median value. The range of the distribution is 17 weeks, but the interquartile range is only 2 weeks. Any gestational period less than 35 weeks or less is considered unusual, with outliers at 25, 26, 27, 28, 29, 30, 31, 32, 33 and 34 weeks.

Exercise 1G

Constructing a five-number summary from a dot or stem plot

1 Construct a five-number summary for the dot plot opposite.



2 Construct a five-number summary for the stem plot opposite.

Key: 13|6 = 136

13	6 7
14	3 6 8 8 9
15	2 5 8 8 8
16	4 5 5 6 7 9
17	8 8 9
18	2 9

ISBN 978-1-009-11041-9

Constructing a boxplot from a five-number summary

Example 22

3 Use the following five-number summaries to construct boxplots.

a Min = 1, $Q_1 = 4$, $M = 8$, $Q_3 = 13.5$, Max = 24

b Min = 136, $Q_1 = 148$, $M = 158$, $Q_3 = 169$, Max = 189

4 **a** Construct a boxplot from the following five-number summary:

Min = 10, $Q_1 = 22$, $M = 40$, $Q_3 = 70$, Max = 70

b Explain why the box has no upper whisker.

5 University participation rates (%) in 21 countries are listed below.

3	3	7	8	9	12	13	15	17	20	21
22	25	26	26	26	27	30	36	37	55	

a Show that the five number summary for this data is:
Min =3, $Q_1 = 10.5$, $M = 21$, $Q_3 = 26.5$, Max = 55

b Show that the upper fence is equal to 50.5.

c Explain why this boxplot will show at least one outlier.

d Construct a boxplot showing the outlier.

6 The five-number summary for a data set is:

Min = 14 $Q_1 = 40$ $M = 55$ $Q_3 = 62$ Max = 99

a Determine the values of the upper and lower fences.

b The smallest three values in the data set are 6, 18, 34 and the largest are 90, 94, 99. Which of these are outliers?

Constructing a boxplot using a CAS calculator

7 The reaction times (in milliseconds) of 18 people are listed below.

38 36 35 35 43 46 42 64 40 48 35 34 40 44 30 25 39 31

a Use a CAS calculator to construct a boxplot with outliers for the data. Name the variable *rtime*.

b Use the boxplot to construct a five-number summary. Identify the outlier.

Reading values from a boxplot

Example 23

8 For each of the boxplots below, estimate the values of:

i the median, M

ii the quartiles Q_1 and Q_3

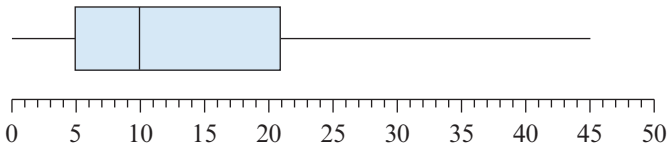
iii the interquartile range, IQR

iv the minimum and maximum values

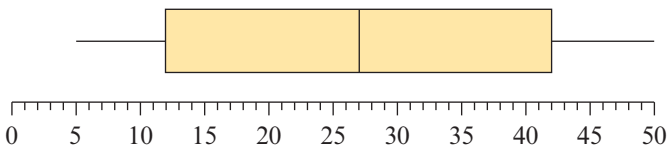
v the values of possible outliers.

ISBN 978-1-009-11041-9

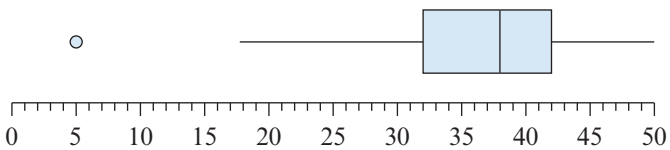
a



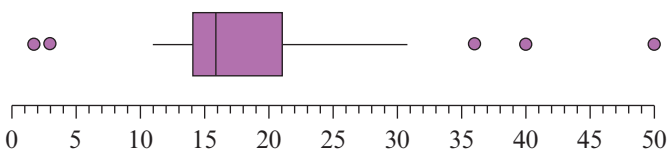
b



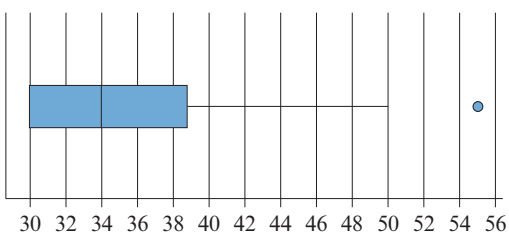
c



d



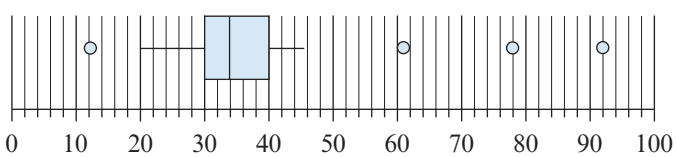
e



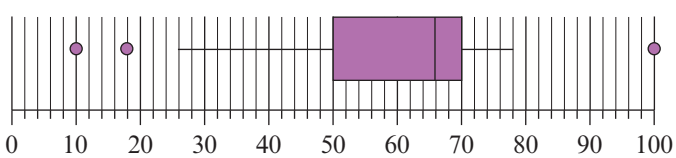
9 For the boxplots below, determine the location of:

i the upper fence **ii** the lower fence.

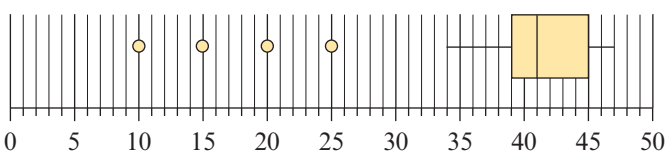
a



b



10 **a** Determine the lower fence for the boxplot opposite.



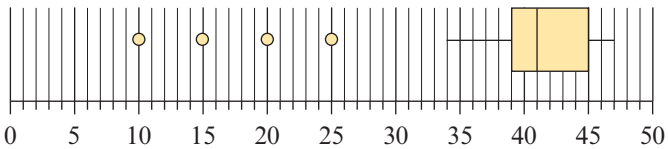
ISBN 978-1-009-11041-9

b When the data were originally entered, a value of 31 was incorrectly entered as 35. Would the 31 be shown as an outlier when the error is corrected? Explain your answer.

Reading percentages from a boxplot

Example 24

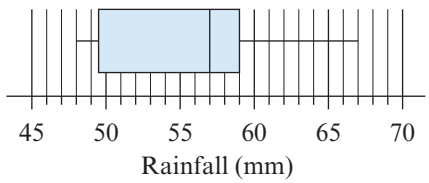
11 Use the boxplot opposite to estimate the percentage of values that are:



a less than 39 **b** less than 45 **c** greater than 45

d between 39 and 45 **e** between 5 and 45.

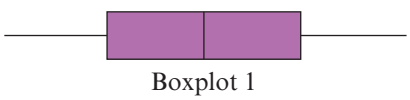
12 The boxplot displays the monthly rainfall (in mm) for 12 months.
Use the boxplot to estimate the percentage of months in which the monthly rainfall was:

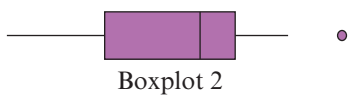


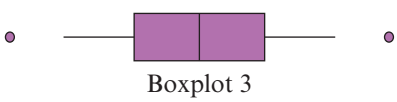
a greater than 59 mm **b** less than 49.5 mm **c** between 49.5 and 59 mm

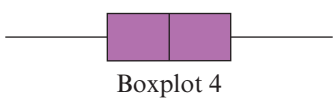
d between 57 and 59 mm **e** less than 59 mm **f** between 57 and 70 mm.

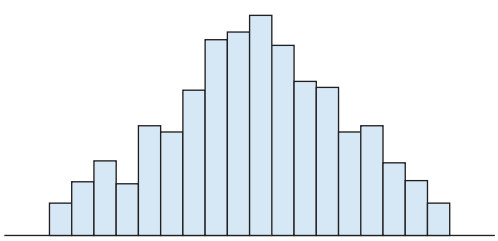
13 Match these boxplots with their histograms.

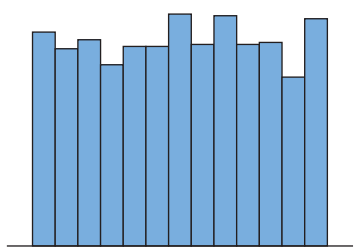

Boxplot 1

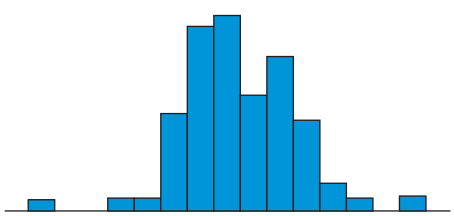

Boxplot 2

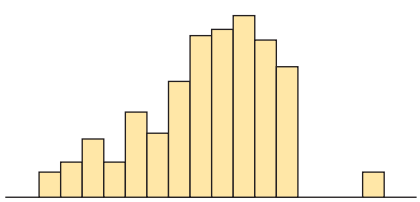

Boxplot 3


Boxplot 4


Histogram A


Histogram B


Histogram C


Histogram D

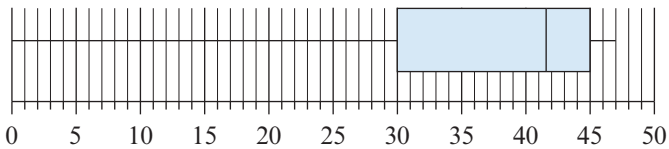
ISBN 978-1-009-11041-9

Describing the features of a distribution from a boxplot

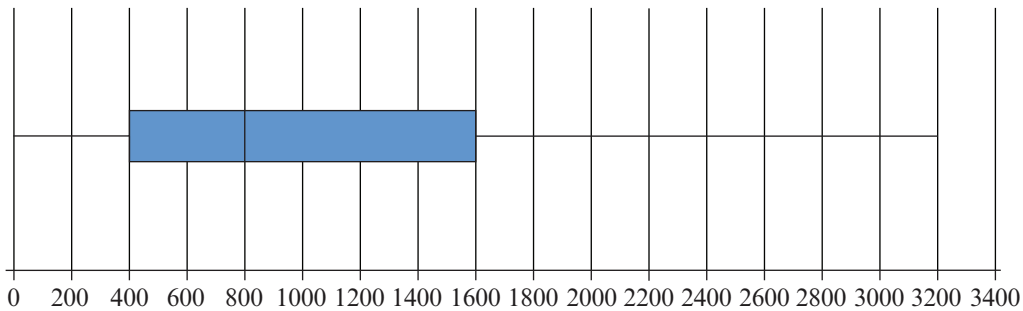
Example 25

14 Describe the distributions represented by the following boxplots in terms of shape, centre, spread. Give appropriate values.

a



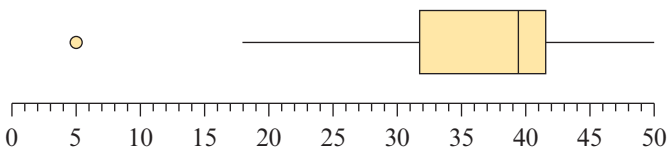
b



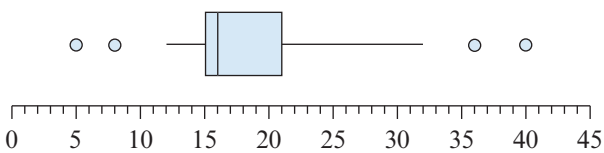
Example 26

15 Describe the distributions represented by the following boxplots in terms of shape, centre, spread and outliers (if any). Give appropriate values.

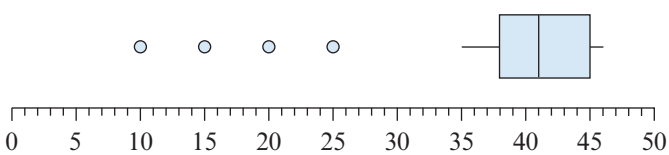
a



b



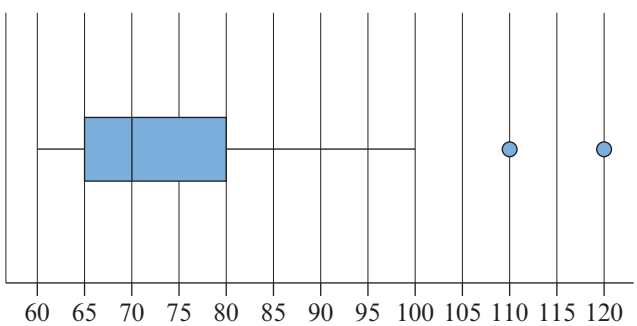
c



Using a boxplot to answer statistical questions

Example 27

16 Taj recorded his travel time to university (in minutes) each day for 60 days, and summarised the data in the following boxplot. Write a brief report describing the distribution of his travel time.



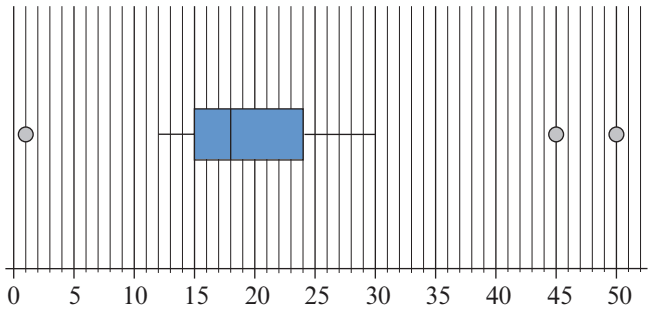
ISBN 978-1-009-11041-9

Cambridge University Press

Exam 1 style questions

Use the following information to answer questions 17 and 18

The boxplot below shows the distribution of marks scored by 200 students in a test.



17 The percentage of students who scored more than 24 marks is closest to:

A 15% **B** 25% **C** 50% **D** 75% **E** 100%

18 The five-number summary for the test scores is closest to:

A 1, 15, 18, 24, 50
B 12, 15, 18, 24, 30
C 1, 12, 15, 18, 24
D 12, 15, 18, 24, 50
E 12, 15, 24, 30, 50

Use the following information to answer questions 19 and 20

The weights (in gm) of 159 fish were measured, and the table gives the mean and the five-number summary for this data.

mean	398
minimum	20
first quartile (Q_1)	120
median (M)	273
third quartile (Q_3)	650
maximum	1650

19 The shape of the distribution is best described as

A approximately symmetric **B** positively skewed
C symmetrically skewed **D** uniform
E negatively skewed

20 The largest five values in the data set are 1100gm, 1250gm, 1550gm, 1600gm and 1650gm. Which of these are outliers?

ISBN 978-1-009-11041-9

A {1100, 1250, 1550, 1600,1650}
B {1650}
C {1250, 1550, 1600,1650}
D {1550, 1600,1650}
E {1600,1650}

1H The normal distribution and the 68–95–99.7% rule

Learning intentions

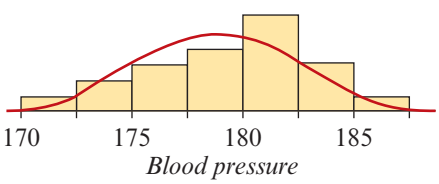
- To be able to introduce the normal model for bell-shaped distributions.
- To be able to use the 68 -95 - 99.7% rule to estimate percentages and give meaning to the standard deviation.
- To be able to calculate standardised scores and use them to compare values across distributions.

We know that the interquartile range is the spread of the middle 50% of the data set. Can we find some similar way in which to interpret the standard deviation?

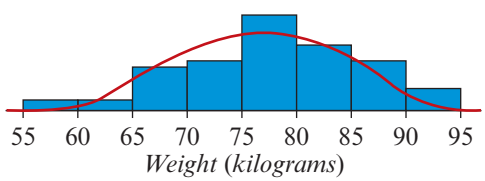
It turns out we can, but we need to restrict ourselves to symmetric distributions that have an approximate bell shape. Again, while this may sound very restrictive, many of the data distributions we work with in statistics (but not all) can be well approximated by this type of distribution. In fact, it is so common that it is called the **normal distribution**.

The normal distribution

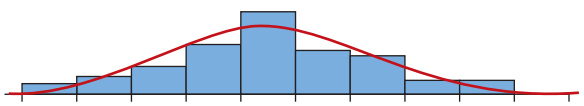
Many data sets that arise in practice are roughly symmetrical and have approximate bell shapes, as shown in the four examples below.



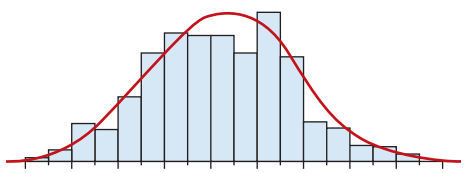
Blood pressure



Weight (kilograms)



Sepal width of flowers (centimetres)



Hand span measurements (centimetres)

Data distributions that are bell-shaped can be modelled by a normal distribution.

The 68–95–99.7% rule

In normal distributions, the percentage of observations that lie within a certain number of standard deviations of the mean can always be determined. In particular, we are interested in the percentage of observations that lie within one, two or three standard deviations of the mean. This gives rise to what is known as the **68–95–99.7% rule**.

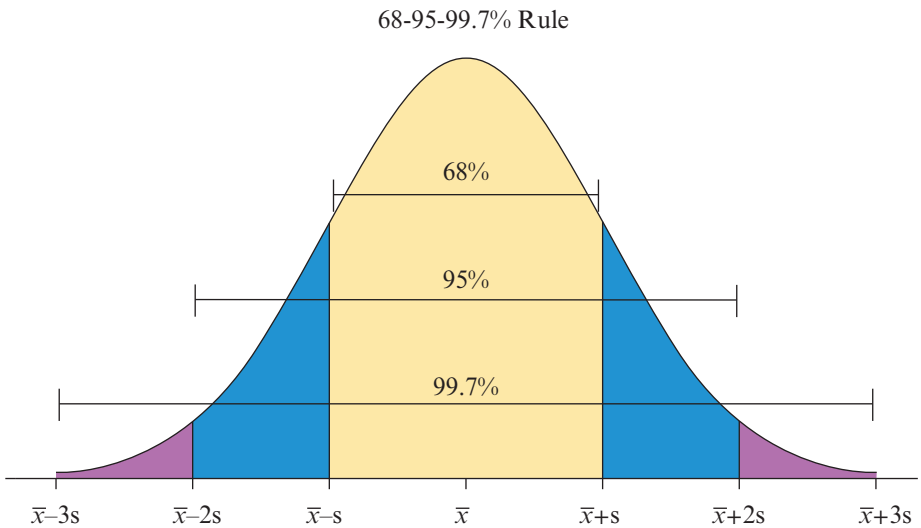
ISBN 978-1-009-11041-9

The 68–95–99.7% rule

For a normal distribution, approximately:

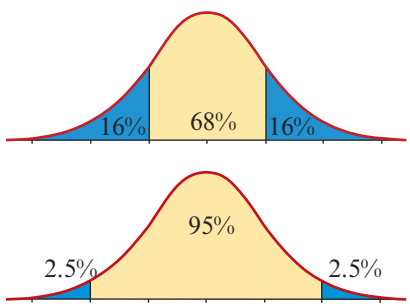
- *68%* of the observations lie within **one** standard deviation of the mean, that is in the interval $(\bar{x} - s, \bar{x} + s)$.
- *95%* of the observations lie within **two** standard deviations of the mean, that is in the interval $(\bar{x} - 2s, \bar{x} + 2s)$.
- *99.7%* of the observations lie within **three** standard deviations of the mean, that is in the interval $(\bar{x} - 3s, \bar{x} + 3s)$.

To give you an understanding of what this rule means in practice, it is helpful to view this rule graphically.



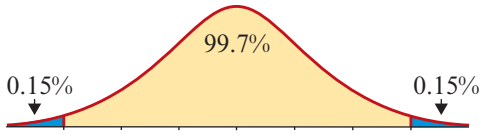
Since the normal distribution is symmetrical, and 100% of the observations are within the normal curve, we can use the 68–95–99.7% rule to allocate percentages to the tails of the distribution in each instance.

- Since around *68%* of the data values will lie within one standard deviation (SD) of the mean, then we can also say that around 16% of values lie in each of the tails.
- Since around 95% of the data values will lie within two standard deviations of the mean, then we can also say that around 2.5% of values lie in each of the tails.

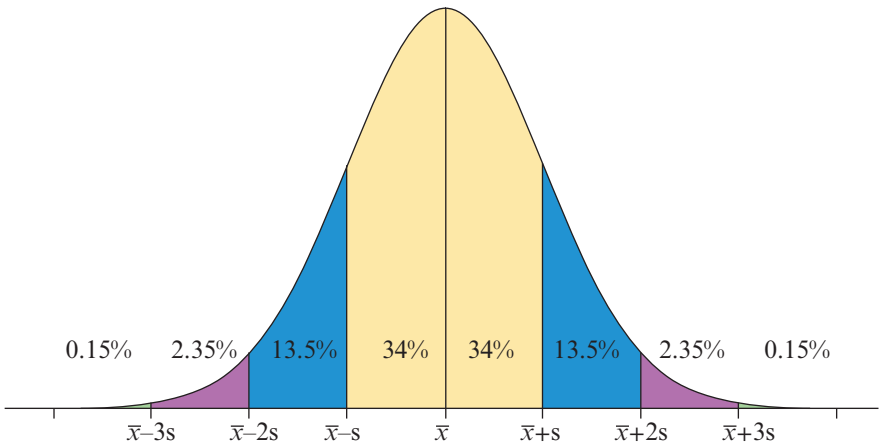


ISBN 978-1-009-11041-9

- Since around 99.7% of the data values will lie within three standard deviations of the mean we can also say that around 0.15% of values lie in each of the tails.



Putting together all of this information, we can then allocate percentages of the data which fall into each section of the normal curve, as shown in the following diagram:



Example 28 Applying the 68–95–99.7% rule

The distribution of delivery times for pizzas made by House of Pizza is approximately normal, with a mean of 25 minutes and a standard deviation of 5 minutes.

a What percentage of pizzas have delivery times of between 15 and 35 minutes?

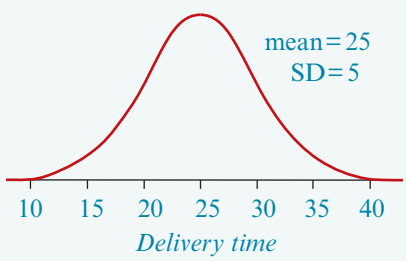
b What percentage of pizzas have delivery times of greater than 30 minutes?

c In 1 month, House of Pizza delivers 2000 pizzas. Approximately many of these pizzas are delivered in less than 10 minutes?

Explanation

a 1 Sketch, scale and label a normal distribution curve with a mean of 25 and a standard deviation of 5.

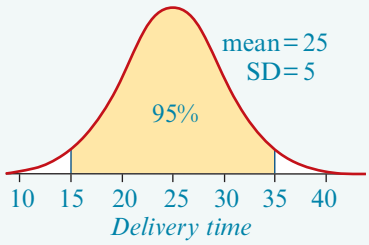
Solution



ISBN 978-1-009-11041-9

2 Shade the region under the normal curve representing delivery times of between 15 and 35 minutes.

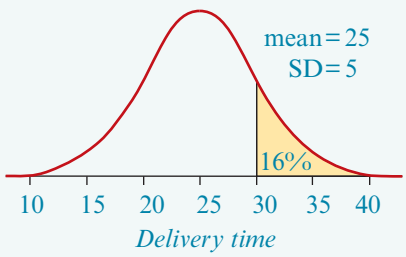
3 Note that delivery times of between 15 and 35 minutes lie within *two* standard deviations of the mean. ($15 = 25 - \mathbf{2} \times 5$ and $35 = 25 + \mathbf{2} \times 5$)



4 95% of values are within two standard deviations of the mean. Use this information to write your answer.

95% of pizzas will have delivery times of between 15 and 35 minutes.

b 1 As before, draw, scale and label a normal distribution curve with a mean of 25 and a standard deviation of 5. Shade the region under the normal curve representing delivery times of greater than 30 minutes.



2 Delivery times of greater than 30 minutes are more than *one* standard deviation above the mean. ($30 = 25 + \mathbf{1} \times 5$)

3 16% of values are more than one standard deviation above the mean. Write your answer.

16% of pizzas will have delivery times of greater than 30 minutes.

c 1 Write down the number of pizzas delivered.

Number = 2000

2 Delivery times of less than 10 minutes are more than *three* standard deviations below the mean. ($10 = 25 - \mathbf{3} \times 5$).

Percentage delivered in less than 10 minutes = 0.15%

3 0.15% of values are more than *three* standard deviations below the mean. Record this.

4 Therefore, the number of pizzas delivered in less than 10 minutes is 0.15% of 2000.

Number of pizzas delivered in less than 10 minutes $\approx$ 0.15% of 2000

$$= \frac{0.15}{100} \times 2000 = 3$$

ISBN 978-1-009-11041-9

Standard scores

The 68–95–99.7% rule makes the standard deviation a natural measuring stick for normally distributed data.

For example, a person who obtained a score of 112 on an IQ test with a mean of 100 and a standard deviation of 15 has an IQ score less than one standard deviation from the mean. Her score is typical of the group as a whole, as it lies well within the middle 68% of scores. In contrast, a person who scores 133 stands out; her score is more than two standard deviations from the mean and this puts her in the top 2.5%.

Because of the additional insight provided by relating the standard deviations to percentages, it is common to transform data into a new set of units that show the number of standard deviations a data value lies from the mean of the distribution. This is called **standardising** and these transformed data values are called **standardised** or ***z*-scores**.

Calculating standardised (z) scores

To obtain a standard score for an actual score, subtract the mean from the score and then divide the result by the standard deviation. That is:

$$\text{standard score} = \frac{\text{actual score} - \text{mean}}{\text{standard deviation}} \quad \text{or} \quad z = \frac{x - \bar{x}}{s}$$

Let us check to see that the formula works.

We already know that an IQ score of 115 is one standard deviation above the mean, so it should have a standard or z-score of 1. Substituting into the formula above we find, as we had predicted, that:

$$z = \frac{115 - 100}{15} = \frac{15}{15} = 1$$

Standard scores can be both positive and negative:

- a positive z-score indicates that the actual score it represents lies above the mean
- a z-score of zero indicates that the actual score is equal to the mean
- a negative z-score indicates that the actual score lies below the mean.

Example 29 Calculating standard scores

The heights of a group of young women have a mean of $\bar{x} = 160$ cm and a standard deviation of $s = 8$ cm. Determine the standard or z-scores of a woman who is:

a 172 cm tall **b** 150 cm tall **c** 160 cm tall.

ISBN 978-1-009-11041-9

Explanation

1 Write down the data value (x), the mean ($\bar{x}$) and the standard deviation (s).

2 Substitute the values into the formula $z = \dfrac{x - \bar{x}}{s}$ and evaluate.

Solution

a $x = 172,\ \bar{x} = 160,\ s = 8$

$$z = \frac{x - \bar{x}}{s} = \frac{172 - 160}{8} = \frac{12}{8} = 1.5$$

b $x = 150,\ \bar{x} = 160,\ s = 8$

$$z = \frac{x - \bar{x}}{s} = \frac{150 - 160}{8} = -\frac{10}{8} = -1.25$$

c $x = 160,\ \bar{x} = 160,\ s = 8$

$$z = \frac{x - \bar{x}}{s} = \frac{160 - 160}{8} = \frac{0}{8} = 0$$

Using standard scores to compare performance

Standard scores are also useful for comparing groups that have different means and/or standard deviations. For example, consider a student who obtained a mark of 75 in Psychology and a mark of 70 in Statistics. In which subject did she do better?

Calculating standard scores

We could take the marks at face value and say that she did better in Psychology because she got a higher mark in that subject. The assumption that underlies such a comparison is that the marks for both subjects have the same distribution with the same mean and standard deviation. However, in this case the two subjects have very different means and standard deviations, as shown in the table above.

Subject	*Mark*	*Mean*	*Standard Deviation*
Psychology	75	65	10
Statistics	70	60	5

If we assume that the *marks* are normally distributed, then standardisation and the 68–95–99.7% rule give us a way of resolving this issue.

Let us standardise the marks.

$$\text{Psychology: standardised mark } z = \frac{75 - 65}{10} = 1$$

$$\text{Statistics: standardised mark } z = \frac{70 - 60}{5} = 2$$

What do we see? The student obtained a higher score for Psychology than for Statistics. However, relative to her classmates she did better in Statistics.

- Her mark of 70 in Statistics is equivalent to a z-score of 2. This means that her mark was two standard deviations above the mean, placing her in the top 2.5% of students.
- Her mark of 75 for Psychology is equivalent to a z-score of 1. This means that her mark was only one standard deviation above the mean, placing her in the top 16% of students. This is a good performance, but not as good as for statistics.

ISBN 978-1-009-11041-9

Cambridge University Press

Example 30 Using standardised scores to make comparisons

Another student studying the same two subjects obtained a mark of 55 for both Psychology and Statistics. Does this mean that she performed equally well in both subjects? Use standardised marks to help you arrive at your conclusion.

Explanation

1 Write down her mark (x), the mean ($\bar{x}$) and the standard deviation (s) for each subject and compute a standardised score for both subjects.

2 Write down your conclusion.

Solution

Pyschology: $x = 55$, $\bar{x} = 65$, $s = 10$

$$z = \frac{x - \bar{x}}{s} = \frac{55 - 65}{10} = \frac{-10}{10} = -1$$

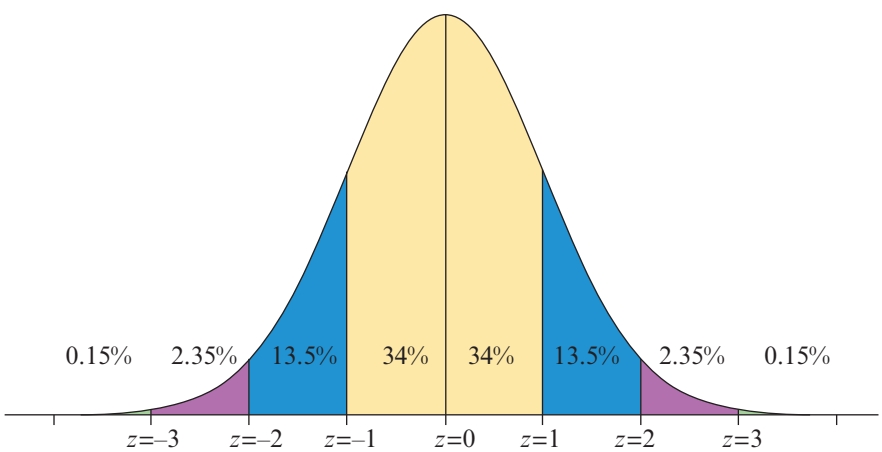
Statistics: $x = 55$, $\bar{x} = 60$, $s = 5$

$$z = \frac{x - \bar{x}}{s} = \frac{55 - 60}{5} = \frac{-5}{5} = -1$$

Yes, her standardised score, $z = -1$, was the same for both subjects. In both subjects she finished in the bottom 16%.

Using standard scores to determine percentages

Now we know how to determine standard scores we can revisit the diagram showing the percentages associated with each section of the normal curve, replacing the horizontal scale with the values of Z (the standard normal scores).



Example 31 Using standard scores to determine percentages

Suppose the weight of a certain species of bird is normally distributed with a mean of 42 grams with a standard deviation of 3 grams.

a If a bird selected at random from this population has a standardised weight of $z = -1$, what percentage of birds in this population weigh more than this bird?

b Approximately what percentage of birds would weigh between 39 and 48 grams?

ISBN 978-1-009-11041-9

Explanation	Solution
a Locate $z = -1$ on the graph above.	We can see that the percentage of the distribution below $z = -1$ is 16%, so the percentage above $z = -1$ is 84%.
b 1 Substitute the values of x into the formula $z = \frac{x - \bar{x}}{s}$ and evaluate.	$\bar{x} = 42, s = 3$ $x = 39 \quad z = \frac{x - \bar{x}}{s} = \frac{39 - 42}{3} = -1$ $x = 48 \quad z = \frac{x - \bar{x}}{s} = \frac{48 - 42}{3} = 2$
2 Locate $z = -1$ and $z = 2$ on the graph above, and determine the percentage of the distribution between these values.	The percentage of the distribution between $z = -1$ and $z = 2$ is 81.5%.

Converting standard scores into actual scores

Having learnt how to calculate standard scores, you also need to be able to convert a standardised score back into an actual score. The rule for converting a standardised score into an actual score is given below.

Converting standardised scores into actual scores

By making the actual score the subject of the rule for calculating standard scores, we arrive at:

actual score = mean + standard score × standard deviation or $x = \bar{x} + z \times s$

Example 32 Converting standard scores into actual scores

A class test (out of 50) has a mean mark of $\bar{x} = 34$ and a standard deviation of $s = 4$. Joe's standardised test mark was $z = -1.5$. What was Joe's actual mark?

Explanation	Solution
1 Write down mean ($\bar{x}$), the standard deviation (s) and Joe's standardised score (z).	$\bar{x} = 34, s = 4, z = -1.5$
2 Write down the rule for calculating the actual score and substitute these values into the rule.	$x = \bar{x} + z \times s$ $= 34 + (-1.5) \times 4 = 28$ Joe's actual mark was 28.

If we know something about the percentages associated with a normal distribution, we can use this information to find the values of the mean, or standard deviation, or both.

In the following example we are given the value of the mean, and one percentage associated with the distribution. From this we can determine the value of the standard deviation.

ISBN 978-1-009-11041-9

Cambridge University Press

Example 33 Finding the value of the standard deviation given the mean and one percentage

Suppose the heights of red flowering gum trees have a mean of 10.2 metres, and 2.5% of these trees grow to more than 11.4 metres tall. Assuming that the heights of these trees are approximately normally distributed, what is the standard deviation of the height of the red flowering gum trees?

Explanation	Solution
1 Since 2.5% of the trees are taller than 11.4 metres, this height corresponds to a z-score of 2.	$\bar{x} = 10.2, z = 2$
2 Write down the rule for calculating the actual score and substitute these values into the rule.	$x = \bar{x} + z \times s$ $11.4 = 10.2 + 2 \times s$
3 Solve this equation for s.	$2 \times s = 1.2$ $s = 0.6$ metres

Example 34 Finding the value of the standard deviation given the mean and two percentages

The marks scored in an examination are known to be approximately normally distributed. If 16% of students score more than 80 marks, and 2.5% of students score less than 20 marks, estimate the mean and standard deviation of this distribution.

Explanation	Solution
1 Since 2.5% of the students score less than 20, this value corresponds to a z-score of -2.	$\bar{x} - 2 \times s = 20$ (1)
2 Since 16% of the students score more than 20, this value corresponds to a z-score of 1.	$\bar{x} + 1 \times s = 80$ (2)
3 To solve these equations for $\bar{x}$ and s, subtract equation (2) from equation (1).	(1)-(2) $3 \times s = 60$ Hence $s = 20$
4 To find $\bar{x}$ substitute the value of s in equation 1.	$\bar{x} - 2 \times 20 = 20$ Hence $\bar{x} = 60$

ISBN 978-1-009-11041-9

Skill-sheet

Exercise 1H

The 68–95–99.7% rule

Example 28

1 The blood pressure readings for executives are approximately normally distributed with a mean systolic blood pressure of 134 and a standard deviation of 20.
Given this information it can be concluded that:

a about 68% of the executives have blood pressures between ☐ and ☐
b about 95% of the executives have blood pressures between ☐ and ☐
c about 99.7% of the executives have blood pressures between ☐ and ☐
d about 16% of the executives have blood pressures above ☐
e about 2.5% of the executives have blood pressures below ☐
f about 0.15% of the executives have blood pressures below ☐
g about 50% of the executives have blood pressures above ☐.

2 The weight of a bag of 10 blood plums picked at U-Pick Orchard is normally distributed with a mean of 1.88 kg and a standard deviation of 0.2 kg.
Given this information the percentage of the bags of 10 plums that weigh:

a between 1.68 and 2.08 kg is approximately ☐ %
b between 1.28 and 2.48 kg is approximately ☐ %
c more than 2.08 kg is approximately ☐ %
d more than 2.28 kg is approximately ☐ %
e less than 1.28 kg is approximately ☐ %
f more than 1.88 kg is approximately ☐ %.

3 The distribution of times taken for walkers to complete a circuit in a park is normal, with a mean time of 14 minutes and a standard deviation of 3 minutes.

a What percentage of walkers complete the circuit in:
i more than 11 minutes? **ii** less than 14 minutes?
iii between 14 and 20 minutes?

b In a week, 1000 walkers complete the circuit. Approximately how many will take less than 8 minutes?

4 The distribution of heights of 19-year-old women is approximately normal, with a mean of 170 cm and a standard deviation of 5 cm.

a What percentage of these women have heights:
i between 155 and 185 cm? **ii** greater than 180 cm?
iii between 160 and 175 cm?

b In a sample of 5000 of these women, approximately how many have heights greater than 175 cm?

ISBN 978-1-009-11041-9

5 The distribution of resting pulse rates of a sample of 2000 20-year-old men is approximately normal, with a mean of 66 beats/minute and a standard deviation of 4 beats/minute.

a What percentage of these men have pulse rates of:

i higher than 66? **ii** between 66 and 70? **iii** between 62 and 74?

b Approximately how many of this sample of 2000 men have pulse rates between 54 and 78 beats/minute?

Calculating standard scores

Example 29

6 A set of scores has a mean of 100 and a standard deviation of 20. Calculate standardised scores for each of the following test scores:

a 120 **b** 140 **c** 80 **d** 100 **e** 40 **f** 110

7 A set of scores has a mean of 30 and a standard deviation of 7. Calculate standardised scores for each of the following test scores. Give your answers to one decimal place.

a 37 **b** 23 **c** 40 **d** 20

Applying standardised scores

Example 30

8 The table below contains the scores a student obtained in a practice test for each of their VCE subjects. Also shown are the mean and standard deviation for each subject.

Subject	*Mark*	*Mean*	*Standard deviation*
English	69	60	4
Biology	75	60	5
Chemistry	55	55	6
Further Maths	55	44	10
Psychology	73	82	4

a Calculate the standardised score for each subject.

b Use the standardised score to rate the student's performance in each subject, assuming a normal distribution of marks and using the 68–95–99.7% rule.

Using standardised scores to determine percentages

Example 31

9 The volume of soft drink in a small can is normally distributed with a mean of 300 mL and a standard deviation of 2 mL.

a If a can selected at random from this population has a standardised volume of $z = 2$, what percentage of cans in this population contain more soft drink than this can?

b Approximately what percentage of cans contain between between 302 mL and 306 mL?

ISBN 978-1-009-11041-9

10 To be considered for a special training program applicants are required to sit for an aptitude test. Suppose that 2000 people sit for the test, and their scores on the aptitude test are approximately normally distributed with a mean of 45 and a standard deviation of 2. People who score more than 49 are selected for the special training program. People who are not chosen for the training program, but score more than 47, are invited to resit the aptitude test at a later date.

a What percentage of people who sat for the test are eligible for the training program?

b Approximately how many people would be invited to resit the aptitude test at a later date?

Calculating actual scores from standardised scores

Example 32

11 A set of scores has a mean of 100 and a standard deviation of 20.
Calculate the actual score if the standardised score was:

a 1 **b** 0.8 **c** 2.1 **d** 0 **e** −1.4 **f** −2.5

Find the values of the mean and standard deviation

Example 33

12 The mean salary for retail assistants is $27/hr. If 2.5% of retail assistants earn more than $30/hr, what is the standard deviation of the salary for retail assistants? Assume that the salaries are approximately normally distributed.

13 The weights of bananas from a certain grower are approximately normally distributed. If the standard deviation of the weight of these bananas is 5 g, and 16% of the bananas weigh less than 96 g, what is the mean weight of the bananas?

Example 34

14 The birth weights of babies are known to be approximately normally distributed. If 16% of babies weigh more than 4.0 kg, and 0.15% of babies weigh more than 5.0 kg, estimate the mean and standard deviation of this distribution. Give your answers to one decimal place.

15 The marks scored in an examination are known to be approximately normally distributed. If 99.7% of students score between 43 and 89 marks, estimate the mean and standard deviation of this distribution. Give your answers to one decimal place.

Applications

16 The body weights of a large group of 14-year-old girls have a mean of 54 kg and a standard deviation of 10.0 kg.

a Kate weighs 56 kg. Determine her standardised weight.

b Lani has a standardised weight of –0.75. Determine her actual weight.

c Find:

i percentage of these girls who weigh more than 74 kg

ii percentage of these girls who weigh between 54 and 64 kg

ISBN 978-1-009-11041-9

Cambridge University Press

iii percentage of these girls who have standardised weights less than –1

iv percentage of these girls who have standardised weights greater than –2.

17 Suppose that IQ scores are normally distributed with mean of 100 and standard deviation of 15.

a What percentage of people have an an IQ:

i greater than 115?

ii less than 70?

b To be allowed to join an elite club, a potential member must have an IQ in the top 2.5% of the population. What IQ score would be necessary to join this club?

c One student has a standardised score of 2.2. What is their actual score?

18 The heights of women are normally distributed with a mean of 160 and a standard deviation of 8.

a What percentage of women would be:

i taller than 152 cm?

ii shorter than 176 cm?

b What height would put a woman among the tallest 0.15% of the population?

c What height would put a woman among the shortest 2.5% of the population?

d One woman has a standardised height of -1.2. What is her actual height? Give your answer to one decimal place.

Exam 1 style questions

Use the following information to answer questions 19 - 22

A total of 16,000 students sat for a statewide mathematics exam. Their results are are normally distributed with mean 50 and standard deviation 7.

19 The percentage of students in the state who scored more than 71 marks is closest to:

A 0.15% **B** 2.5% **C** 5% **D** 15% **E** 0.3%

20 The top 2.5% of the state are to be awarded a distinction. What would be the lowest mark required to gain a distinction in this exam?

A 36 **B** 43 **C** 57 **D** 64 **E** 71

21 Approximately how many students gained a mark of 57 or more?

A 400 **B** 800 **C** 2560 **D** 5120 **E** 15200

22 Approximately how many students gained a mark between 43 and 64?

A 10888 **B** 11120 **C** 13040 **D** 13440 **E** 15200

ISBN 978-1-009-11041-9

Cambridge University Press

23 The table below shows Miller's swimming times (in seconds) for 50 metres in each of butterfly, backstroke, breaststroke and freestyle. Also shown are the mean and standard deviation of the times recorded for all of the swimmers in his swimming club. In how many of these swimming styles is he in the fastest 2.5% of swimmers at his swimming club?

Style	*Miller's time*	*Mean*	*Standard deviation*
Butterfly	38.8	46.2	3.2
Breaststroke	51.4	55.1	4.1
Backstroke	53.5	48.3	2.5
Freestyle	33.3	38.2	2.3

A 0 **B** 1 **C** 2 **D** 3 **E** 4

1I Populations and samples

This material is available in the Interactive Textbook.

ISBN 978-1-009-11041-9

Review

Key ideas and chapter summary

Assign-ment

Univariate data

Univariate data are generated when each observation involves recording information about a single variable.

Types of data

Data can be classified as numerical or categorical.

Categorical variables

Categorical variables are used to represent characteristics of individuals. Categorical variables come in two types: nominal and ordinal. **Nominal variables** generate data values that can only be used by name, e.g. eye colour. **Ordinal variables** generate data values that can be used to both name and order, e.g. house number.

Numerical variables

Numerical variables have data values which are quantities. Numerical variables come in two types: discrete and continuous. **Discrete variables** are those which may take on only a countable number of distinct values such as 0 1 2 3 4 . . . and are often associated with counting. **Continuous variables** are ones which take an infinite number of possible values, and are often associated with measuring rather than counting.

Frequency table

A **frequency table** lists the values a variable takes, along with how often (frequently) each value occurs. Frequency can be recorded as:

- the number of times a value occurs – e.g. the number of Year 12 students in the data set is 32.
- the percentage of times a value occurs – e.g. the percentage of Year 12 in the data set is 45.5%.

Bar chart

Bar charts are used to display frequency distribution of categorical data. Each value of the variable is represented by a bar showing the frequency, or the percentage frequency.

Segmented bar chart

A **segmented bar chart** is like a bar chart, but the bars are stacked one on top of another to give a single bar with several segments.

Mode, modal category

The **mode** (or modal interval) is the value of a variable (or the interval) that occurs most frequently.

Histogram

A **histogram** uses columns to display the frequency distribution of a numerical variable: suitable for medium to large-sized data sets.

ISBN 978-1-009-11041-9

Describing the distribution of a numerical variable

The distribution of a numerical variable can be described in terms of:

- shape: symmetric or skewed (positive or negative)
- outliers: values that seem unusually small or large.
- centre: the median or mean.
- spread: the IQR , range or the standard deviation.

Dot plot

A **dot plot** consists of a number line with each data point marked by a dot. A dot plot is particularly suitable for displaying a small data set of discrete numerical data.

Stem plot

The **stem plot** is particularly suitable for displaying a small to medium sized data sets of numerical data.It shows each data value separated into two parts: the leading digits, which make up the stem of the number, and its last digit, which is called the leaf.

Log scales

Log scales can be used to transform a skewed histogram to symmetry.

Summary statistics

Summary statistics are numerical values for special features of a data distribution such as centre and spread.

Mean

The **mean ($\bar{x}$)** is a summary statistic that can be used to locate the centre of a symmetric distribution. The value of the mean is determined from the formula: $\bar{x} = \dfrac{\sum x}{n}$

Range

The **range *(R)*** is the difference between the smallest and the largest data values. It is the simplest measure of spread.

Standard deviation

The **standard deviation (s)** is a summary statistic that measures the spread of the data values around the mean. The value of the standard deviation is determined from the formula:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Median

The **median *(M)*** is a summary statistic that can be used to locate the centre of a distribution. It is the midpoint of a distribution, so that 50% of the data values are less than this value and 50% are more. It is sometimes denoted as Q_2

Quartiles

Quartiles are summary statistics that divide an ordered data set into four equal groups.

ISBN 978-1-009-11041-9

Cambridge University Press

Review

Interquartile range

The **interquartile range (*IQR*)** gives the spread of the middle 50% of data values in an ordered data set. It is found by evaluating $IQR = Q_3 - Q_1$

Five-number summary

The median, the first quartile, the third quartile, along with the minimum and the maximum values in a data set, are known as a **five-number summary**.

Outliers

Outliers are data values that appear to stand out from the rest of the data set. They are values that are less than the **lower fence** or more than the **upper fence**.

Lower and upper fences

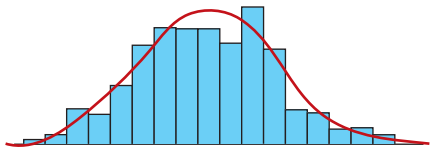
The **lower fence** is equal to $Q_1 - 1.5 \times IQR$.
The **upper fence** is equal to $Q_3 + 1.5 \times IQR$.

Boxplot

A **boxplot** is a visual display of a five-number summary with adjustments made to display outliers separately when they are present.

The normal distribution

Data distributions that have a bell shape can be modelled by a **normal distribution.**



The 68-95-99.7%rule

For a data distribution which is approximately normally distributed approximately:

- 68% of the data will lie within one standard deviation of the mean.
- 95% of the data will lie within two standard deviations of the mean.
- 99.7% of the data will lie within three standard deviations of the mean.

Standardised scores

The value of the standard score gives the distance and direction of a data value from the mean in terms of standard deviations.
The rule for calculating a **standardised score** is:

$$\text{standardised score} = \frac{\text{actual score} - \text{mean}}{\text{standard deviation}}$$

ISBN 978-1-009-11041-9

Skills checklist

Download this checklist from the Interactive Textbook, then print it and fill it out to check your skills.

1A	1	I can identify types of data.	☐
		See Example 1, and Exercise 1A Question 1	
1B	2	I can construct a frequency table for categorical data.	☐
		See Example 2, and Exercise 1B Question 1	
1B	3	I can construct a bar chart from a frequency table.	☐
		See Example 3, and Exercise 1B Question 2	
1B	4	I can construct a percentage segmented bar chart from a frequency table.	☐
		See Example 4, and Exercise 1B Question 4	
1B	5	I can describe the distribution of a categorical variable.	☐
		See Example 5, and Exercise 1B Question 6	
1C	6	I can construct a frequency table for discrete numerical data.	☐
		See Example 6, and Exercise 1C Question 1	
1C	7	I can construct a grouped frequency table.	☐
		See Example 7, and Exercise 1C Question 3	
1C	8	I can construct a histogram from a grouped frequency table.	☐
		See Example 8, and Exercise 1C Question 4	
1C	9	I can describe the features of a distribution from a histogram.	☐
		See Example 9, and Exercise 1C Question 9	
1D	10	I can construct a dot plot.	☐
		See Example 10, and Exercise 1D Question 1	
1D	11	I can construct a stem plot.	☐
		See Example 11, and Exercise 1D Question 6	
1E	12	I can use a CAS calculator to find logs.	☐
		See Example 12, and Exercise 1E Question 1	

ISBN 978-1-009-11041-9

Cambridge University Press

Review

1E	13	I can interpret a histogram with a log scale. ☐
		See Example 13, and Exercise 1E Question 4
1F	14	I can find the median value in a data set. ☐
		See Example 14, and Exercise 1F Question 1
1F	15	I can find the median value from a dot plot. ☐
		See Example 15, and Exercise 1F Question 3
1F	16	I can find the median value from a stem plot. ☐
		See Example 16, and Exercise 1F Question 4
1F	17	I can find the range from a stem plot. ☐
		See Example 17, and Exercise 1F Question 4
1F	18	I can find the *IQR* from a stem plot when *n* is even. ☐
		See Example 18, and Exercise 1F Question 5
1F	19	I can find the *IQR* from a stem plot when *n* is odd. ☐
		See Example 19, and Exercise 1F Question 8
1F	20	I can calculate the mean using the formula. ☐
		See Example 21, and Exercise 1F Question 10
1F	21	I can calculate the mean and standard deviation using the CAS calculator. ☐
		See CAS 3, and Exercise 1F Question 12
1G	22	I can construct a boxplot from a five number summary. ☐
		See Example 22, and Exercise 1G Question 3
1G	23	I can construct a boxplot with outliers from data using the CAS calculator. ☐
		See CAS 4, and Exercise 1G Question 7
1G	24	I can read values from a boxplot. ☐
		See Example 23, and Exercise 1G Question 8
1G	25	I estimate percentages from a boxplot. ☐
		See Example 24, and Exercise 1G Question 11
1G	26	I can use boxplots for describe a distribution without outliers. ☐
		See Example 25, and Exercise 1G Question 14

ISBN 978-1-009-11041-9

Cambridge University Press

1G	**27** I can use boxplots to describe a distribution with outliers.	☐
	See Example 26, and Exercise 1G Question 15	
1G	**28** I can use boxplots to answer statistical questions.	☐
	See Example 27, and Exercise 1G Question 16	
1H	**29** I can apply the 68-95-99.7% rule.	☐
	See Example 28, and Exercise 1H Question 1	
1H	**30** I can calculate standardised scores.	☐
	See Example 29, and Exercise 1H Question 6	
1H	**31** I can use standardised scores to make comparisons.	☐
	See Example 30, and Exercise 1H Question 8	
1H	**32** I can use standard scores to determine percentages.	☐
	See Example 31, and Exercise 1H Question 9	
1H	**33** I can convert standardised scores into actual scores.	☐
	See Example 32, and Exercise 1H Question 11	
1H	**34** I can solve for the values of the mean and standard deviation.	☐
	See Example 33 and 34, Exercise 1H Question 12 and Question 14.	

Multiple-choice questions

The following information relates to Questions 1 and 2.

Data pertaining to the following five variables was collected about secondhand cars:

- *colour*
- *model*
- *number of seats*
- *age* (1 = less than 2 years, 2 = from 2 years to 5 years, 3 = more than 5 years)
- *mileage* (in kilometres)

1 The number of these variables that are discrete numerical is:

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

2 The number of ordinal variables is:

A 0 **B** 1 **C** 2 **D** 3 **E** 4

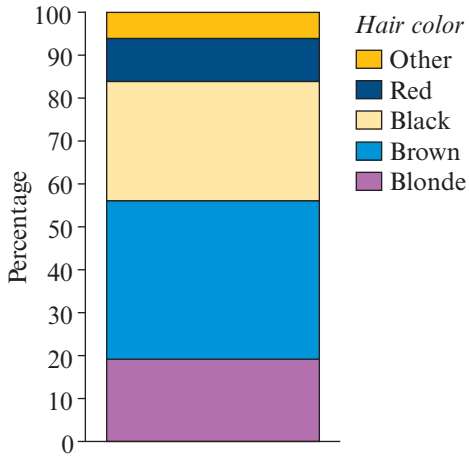
ISBN 978-1-009-11041-9

Cambridge University Press

Review

The following information relates to Questions 3 and 4.

The percentage segmented bar chart shows the distribution of hair colour for 200 students.



3 The number of students with brown hair is closest to:

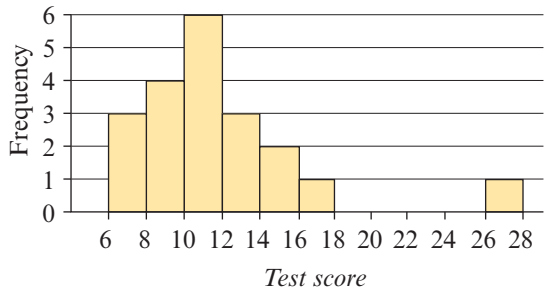
A 4 **B** 34 **C** 57 **D** 72 **E** 114

4 The most common hair colour is:

A black **B** blonde **C** brown **D** red **E** other

Questions 5 to 7 relate to the histogram shown below.

The histogram below displays the test scores of a class of students.



5 The number of students in the class who obtained a test score less than 14 is:

A 4 **B** 10 **C** 14 **D** 16 **E** 28

6 The shape of the histogram is best described as:

A negatively skewed
B positively skewed with an outlier
C approximately symmetric
D approximately symmetric with an outlier
E negatively skewed with an outlier

7 The value of the first quartile could be:

A 5.5 **B** 6.8 **C** 8.9 **D** 10.5 **E** 11.4

8 $\log_{10} 100$ equals:

A 0 **B** 1 **C** 2 **D** 3 **E** 100

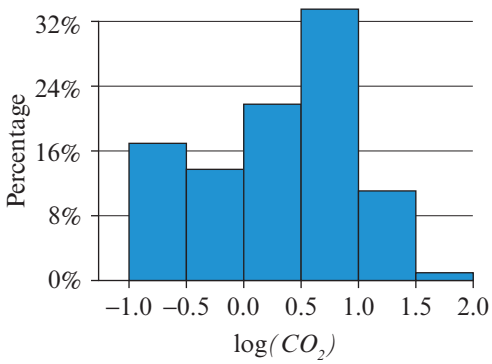
9 Find the number with log equal to 2.314; give the answer to the nearest whole number.

A 2 **B** 21 **C** 206 **D** 231 **E** 20606

ISBN 978-1-009-11041-9

The following information relates to Questions 10 and 11.

The percentage histogram opposite displays the distribution of the log of the annual per capita CO_2 emissions (in tonnes) for 192 countries in 2011.



10 Australia's per capita CO_2 emissions in 2011 were 16.8 tonnes. In which column of the histogram would Australia be located?

A −0.5 to <0.0 **B** 0.0 to <0.5 **C** 0.5 to <1.0 **D** 1.0 to <1.5 **E** 1.5 to < 2.0

11 The percentage of countries with per capita CO_2 emissions of under 10 tonnes is closest to:

A 14% **B** 17% **C** 31% **D** 69% **E** 88%

12 The following is an ordered set of 10 daily maximum temperatures (in degrees Celsius):

22 22 23 24 24 25 26 27 28 29

The five-number summary for these temperatures is:

A 22, 23, 24, 27, 29 **B** 22, 23, 24.5, 27, 29 **C** 22, 24, 24.5, 27, 29
D 22, 23, 24.5, 27.5, 29 **E** 22, 24, 24.5, 27, 29

The following information relates to Questions 13 to 15.

The stem plot opposite displays the distribution of the marks obtained by 25 students.

Key: 1|5 means 15 marks

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

13 The median mark is:

A 20 **B** 27 **C** 28 **D** 29 **E** 30

14 The interquartile quartile range (*IQR*) of the marks is:

A 12 **B** 16.5 **C** 20 **D** 30.5 **E** 31.5

ISBN 978-1-009-11041-9

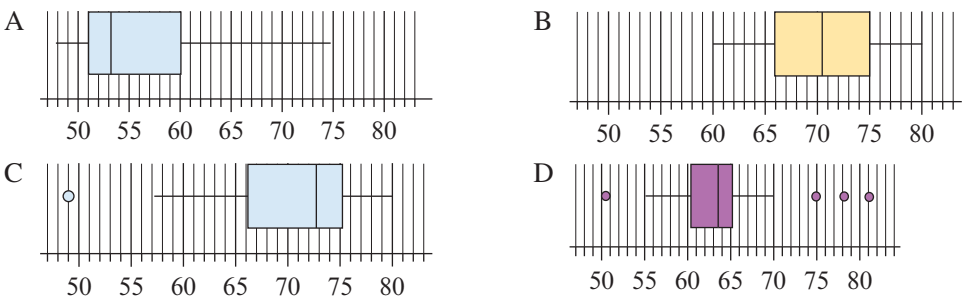
Cambridge University Press

Review

15 The shape of the data distribution displayed by this stem plot is best described as:

A approximately symmetric
B approximately symmetric with an outlier
C negatively skewed with an outlier
D negatively skewed
E positively skewed with an outlier

The following information relates to Questions 16 to 23.



16 The median of boxplot A is closest to:

A 5 **B** 53 **C** 54.5 **D** 55 **E** 60

17 The *IQR* of boxplot B is closest to:

A 9 **B** 20 **C** 25 **D** 65 **E** 75

18 The range of boxplot C is closest to:

A 4 **B** 13 **C** 20 **D** 31 **E** 80

19 The description that best matches boxplot A is:

A symmetric
B symmetric with outliers
C negatively skewed
D positively skewed
E positively skewed with outliers

20 The description that best matches boxplot B is:

A symmetric
B negatively skewed with an outlier
C negatively skewed
D positively skewed
E positively skewed with outliers

21 The description that best matches boxplot D is:

A symmetric
B symmetric with outliers
C negatively skewed
D positively skewed
E positively skewed with outliers

22 For the data represented by boxplot D, the percentage of data values greater than 65 is:

A 2.5% **B** 25% **C** 50% **D** 75% **E** 100%

ISBN 978-1-009-11041-9

23 To be an outlier in boxplot D, a score must be:

A either less than 52.5 or greater than 72.5 **B** greater than 72.5
C either less than 55 or greater than 70 **D** greater than 70
E less than 55

24 The mean ($\bar{x}$) and standard deviation (s) for the following set of test marks

1 1 10 15 16 25 8 10 12

are closest to:

A 7.1, 10.9 **B** 7.5, 10.9 **C** 10.9, 7.1 **D** 10.9, 7.5 **E** 10.8, 7.5

25 It would *not* be appropriate to determine the mean and standard deviation of a group of peoples':

A ages **B** phone numbers **C** heights
D weights **E** family sizes

26 The median is a more appropriate measure of the centre of a distribution than the mean when the distribution is:

A symmetric **B** symmetric with no outliers **C** bell-shaped
D clearly skewed **E** normal

27 The mean length of 10 garden stakes is 180.5 cm, and the standard deviation is 2.9 cm. If the length of each garden stake is reduced by exactly 5 cm, the mean and standard deviation of the 10 stakes will be:

A 175.5 cm and 2.4 cm **B** 180.5 cm and 2.4 cm **C** 175.5 cm and 2.9 cm
D 175.5 cm and 3.4 cm **E** 185.5 and 2.9 cm

28 A student's mark on a test is 50. The mean mark for their class is 55 and the standard deviation is 2.5. Their standard score is:

A −2.5 **B** −2.0 **C** 0 **D** 2 **E** 2.5

The following information relates to Questions 29 to 33.

Each week, a bus company makes 200 trips between two large country towns. The time taken to make a trip between the two towns is approximately normally distributed with a mean of 78 minutes and a standard deviation of 4 minutes.

29 The percentage of trips each week that take 78 minutes or more is:

A 16% **B** 34% **C** 50% **D** 68% **E** 84%

30 The number of trips each week that take between 70 and 82 minutes is approximately:

A 4 **B** 32 **C** 68 **D** 127 **E** 163

ISBN 978-1-009-11041-9

Cambridge University Press

31 A trip that takes 71 minutes has a standardised time (z-score) of:

A −1.75 **B** −1.5 **C** −1.25 **D** 1.5 **E** 1.75

32 A standardised time for a trip is $z = -0.25$. The actual time (in minutes) is:

A 77 **B** 77.25 **C** 77.75 **D** 78.25 **E** 79

33 The time of a bus trip has a standardised time of $z = 2.1$. This time is:

A very much below average **B** just below average **C** around average
D just above average **E** very much above average

34 The table shows the time taken to run one kilometre (in minutes) by three runners. To be invited to join the athletics team the standardised score for their time needs to be no more than than 0.5.

Runner	*Time(mins)*
Albie	7.5
Lincoln	4.9
Wendy	8.0

If the mean running time for one kilometre is 7.0 minutes and the standard deviation is 1.2 minutes, who will be invited to join the athletics team?

A Only Wendy **B** Lincoln and Albie **C** Only Lincoln
D Albie and Wendy **E** All three runners

35 The diameter of bolts produced by a machine is normally distributed. If 2.5% of bolts the bolts have a diameter of more than 4.94 mm, and 0.15% have a diameter less than 4.84 mm, the mean and standard deviation of this distribution in millimetres are closest to:

A mean = 4.88 standard deviation =0.02
B mean = 4.92 standard deviation = 0.01
C mean = 4.90 standard deviation = 0.001
D mean = 4.90 standard deviation = 0.1
E mean = 4.90 standard deviation = 0.02

Written response questions

1 A group of 52 teenagers were asked, 'Do you agree that the use of marijuana should be legalised?' Their responses are summarised in the table.

	Frequency	
Legalise	Number	Percentage
Agree	18	
Disagree	26	
Don't know	8	
Total	52	

a Construct a labelled and scaled frequency bar chart for the data.

b Complete the table by calculating the percentages to one decimal place.

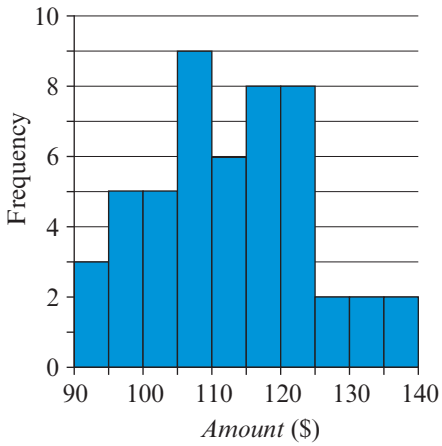
ISBN 978-1-009-11041-9

Cambridge University Press

c Use the percentages to construct a percentage segmented bar chart for the data.

d Write a short report describing the distribution of responses.

2 Students were asked how much they spent on entertainment each month. The results are displayed in the histogram. Use the histogram to answer the following questions.



a How many students:

i were surveyed?

ii spent from \$100 to less than \$105 per month?

b What is the mode?

c How many students spent \$110 or more per month?

d What percentage spent less than \$100 per month?

e **i** What is the shape of the distribution displayed by the histogram?

ii In which interval is the median of the distribution?

iii In which interval is the upper quartile of the distribution (Q_3)?

3 The amount of weight lost in one week by 32 people who participated in a weight loss program was recorded and displayed in the ordered stem plot below.

Weight loss (kg) key: 2|0 represents 2.0

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

a Describe the shape of the distribution.

b Determine the median weight loss. Give your answer to 2 decimal places.

c Find the value of the interquartile range. Give your answer to 2 decimal places.

d What percentage of this group had a weight loss of more than 3.5 kg? Give your answer to 2 decimal places.

e Is the weight loss of 5.0 kg an outlier for this data set? Justify your answer.

4 The systolic blood pressure (measured in mmHg) for a group of 2000 people was measured. The results are summarised in the five-number summary below:
Min = 75, Q_1 = 110, M = 125, Q_3 = 140, Max = 180

a Use the five-number summary to construct a simple boxplot.

ISBN 978-1-009-11041-9

Cambridge University Press

b Indicate on your plot where the lower and upper fences would be, and hence if there would be any outliers.

c Assume that the distribution of systolic blood pressure for this sample of 2000 people is approximately normally distributed, with a mean of 128 mmHg and a standard deviation of 20 mmHg.

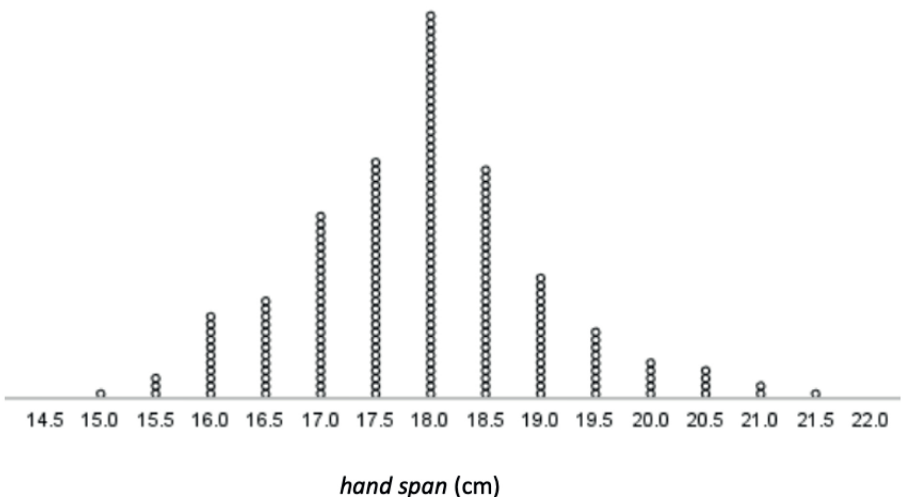
i Approximately what percentage of people of people have a systolic blood pressure between 108 mmHg and 148 mmHg?

ii Suppose a person has a blood pressure three standard deviations below the mean, what would be their actual blood pressure?

iii Of the 2000 people measured, how many could we expect to have a blood pressure three standard deviations below the mean?

iv Of the 2000 people measured, how many actually did have a blood pressure three standard deviations below the mean?

5 The *hand span* in centimetres of 200 women was recorded and displayed in the dot plot below.



a Write down the modal *hand span*, in centimetres, for this group of 200 women.

b The mean *hand span* for this group of 200 women is 17.9 cm, and the standard deviation is 1.1 cm. Use the information in the dot plot to determine the percentage of women in this group who had an actual hand span more than two standard deviations above or below the mean. Round your answer to one decimal place.

c The five-number summary for this sample of hand spans, in centimetres, is given below:

Min =15.0, Q_1 = 17.0, M = 18.0, Q_3 =18.5, Max = 21.5

Use the five-number summary to construct a boxplot showing outliers.

ISBN 978-1-009-11041-9

Cambridge University Press