Chapter 2

# 2

# Investigating and comparing data distributions

- What are categorical and numerical data?
- What is a bar chart and when is it used?
- What is a histogram and when is it used?
- What is are dot and stem plots and when are they used?
- What are the mean, median, range, interquartile range and standard deviation?
- What are the properties of these summary statistics and when is each used?
- How do we construct and interpret boxplots?

ISBN 978-1-107-56755-9

Cambridge University Press

## Introduction

In this information age we increasingly have to interpret data. This data may be presented in charts, diagrams or graphs, or it may simply be lists of words or numbers. There may be a lot of relevant information embodied in the data, but the story it has to tell will not always be immediately obvious. Various statistical procedures are available which will help us extract the relevant information from data sets. In this chapter, we will look at some of the techniques that are used when the data are collected from a single variable and that can help us to answer real world questions.

## 2A Types of data

Consider the following situation. In completing a survey, students are asked to:

- indicate their sex by circling an 'F' for female or an 'M' for male on the form
- indicate their preferred coffee cup size when buying takeaway coffee as 'small', 'medium' or 'large'
- write down the number of brothers they have
- measure and write their hand span in centimetres.

The information collected from four students is displayed in the table below.

Since the answers to each of the questions in the survey will vary from student to student, each question defines a different **variable** namely: *sex*, *coffee size*, *number of brothers* and *hand span*. The values we collect about each of these variables are called **data**.

| Sex | Coffee size | Number of brothers | Hand span (in cm) |
|---|---|---|---|
| M | Large | 0 | 23.6 |
| F | Small | 2 | 19.6 |
| F | Small | 1 | 20.2 |
| M | Large | 1 | 24.0 |

The data in the table fall into two broad types: *categorical* or *numerical*.

### ▶ Categorical data

The data arising from the students responses to the first and second questions in the survey are called **categorical data** because the data values can be used to place the person into one of several groups or categories. However, the properties of the data generated by these two questions differ slightly.

- The question asking students to use an 'M' or 'F' to indicate their sex will prompt a response of either M or F. This identifies the respondent as either male or female but tells us no more. We call this **nominal data** because it simply *names* or *nominates*.
- However, the question with responses 'small', 'medium' and 'large' that indicates the students' preferred coffee size tells us two things. Firstly, it names the coffee size, but secondly it enables us to order the students according to their preferred coffee sizes. We call this **ordinal data** because it enables us to both name and order their responses.

## Numerical data

The data arising from the responses to the third and fourth questions in the survey are called **numerical data** because they have values for which arithmetic operations such as adding and averaging make sense. However, the properties of the data generated by these questions differs slightly.

- The question asking students to write down the number of brothers they have will prompt whole number responses like $0, 1, 2, \ldots$
  Because the data can only take particular numerical values it is called **discrete data**. Discrete data arises in situations where counting is involved. For this reason, discrete data is sometimes called count data.
- In response to the hand span question, students who wrote 24 cm could have an actual hand span of anywhere between 23.5 and 24.4 cm, depending on the accuracy of the measurement and how the student rounded their answer. This is called **continuous data**, because the variable we are measuring, in this case, *hand span*, can take any numerical value within a specified range.
  Continuous data are often generated when measurement is involved. For this reason, continuous data is sometimes called measurement data.

## Types of variables

### Categorical variables

Variables that generate categorical data are called **categorical variables** or, if we need a finer distinction, **nominal** or **ordinal** variables. For example, *sex* is a nominal variable, while *coffee size* is an ordinal variable.

### Numerical variables

Variables that generate numerical data are called numerical variables or, if we need a finer distinction, discrete or continuous variables. For example, *number of brothers* is a **discrete variable**, while *hand span* is a continuous variable.

ISBN 978-1-107-56755-9

Cambridge University Press

## Exercise 2A

### Classifying data

**1** Classify the categorical data arising from people answering the following questions as either nominal or ordinal.

**a** What is your favourite football team?

**b** How often do you exercise? Choose one of ‘never’, ‘once a month’, ‘once a week’, ‘every day’.

**c** Indicate how strongly you agree with ‘alcohol is the major cause of accidents’ by selecting one of ‘strongly agree’, ‘agree’, ‘disagree’, ‘strongly disagree’.

**d** What language will you study next year, ‘French’, ‘Chinese’, ‘Spanish’ or ‘none’?

**2** Classify the data generated in each of the following as categorical or numerical.

**a** Kindergarten pupils bring along their favourite toys, and they are grouped together under the headings ‘dolls’, ‘soft toys’, ‘games’, ‘cars’ and ‘other’.

**b** The number of students on each of 20 school buses are counted.

**c** A group of people each write down their favourite colour.

**d** Each student in a class is weighed in kilograms.

**e** Students are weighed and then classified as ‘light’, ‘average’ or ‘heavy’.

**f** People rate their enthusiasm for a certain rock group as ‘low’, ‘medium’ or ‘high’.

**3** Classify the data generated in each of the following situations as nominal, ordinal or numerical (discrete or continuous).

**a** The different brand names of instant soup sold by a supermarket are recorded.

**b** A group of people are asked to indicate their attitude to capital punishment by selecting a number from 1 to 5, where 1 = strongly disagree, 2 = disagree, 3 = undecided, 4 = agree and 5 = strongly agree.

**c** The number of computers per household was recorded during a census.

### Classifying variables

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

**a** The *number of pages* in a book

**b** The *price* paid 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 *time* between the arrival of successive customers at an ATM

**e** The *number of people* at a football match

# 2B Displaying and describing categorical data distributions

To make sense of data, we first need to organise it into a more manageable form. For categorical data, frequency tables and bar charts are used for this purpose.

## ▶ The frequency table

### Frequency

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:

- **frequency**: the number of times a value occurs
- **percentage frequency**: the percentage of times a value occurs, where:

$$\text{percentage frequency} = \frac{\text{count}}{\text{total}} \times 100\%$$

- frequency **distribution**: a listing of the values a variable takes, along with how frequently each of these values occurs.

### Example 1 Constructing a frequency table for categorical data

Thirty children chose a sandwich, a salad or a pie for lunch, as follows:

> sandwich, salad, salad, pie, sandwich, sandwich, salad, salad, pie, pie, pie, salad, pie, sandwich, salad, pie, salad, pie, sandwich, sandwich, pie, salad, salad, pie, pie, pie, salad, pie, sandwich, pie

Construct a table for the data showing both frequency and percentage frequency.

### Solution

| | Frequency | |
|---|---|---|
| Lunch choice | Number | % |
| Sandwich | 7 | 23.3 |
| Salad | 10 | 33.3 |
| Pie | 13 | 43.3 |
| Total | 30 | 99.9 |

1 Set up a table as shown. The variable *lunch choice* has three categories: 'sandwich', 'salad' and 'pie'.

2 Count the number of children choosing a sandwich, a salad or a pie. Record in the 'Number' column.

3 Add the frequencies to find the total number.

4 Convert the frequencies into percentages and record in the '%' column. For example,

percentage frequency for pies equals $\frac{13}{30} \times 100\% = 43.3\%$

5 Total the percentages and record. Note that the percentages add up to 99.9%, not 100%, because of rounding.

ISBN 978-1-107-56755-9

Cambridge University Press

## Bar charts

When there are a lot of data, a frequency table can be used to summarise the information, but we generally find that a graphical display is also useful. When the data are categorical, the appropriate display is a **bar chart**.

### Bar charts

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 (or percentage)
- the bars are drawn with gaps to indicate that each value is a separate category
- there is one bar for each category.

### Example 2 Constructing a bar and percentage bar chart from a frequency table

Use the frequency table for lunch choice from Example 1 to construct:

**a** a bar chart

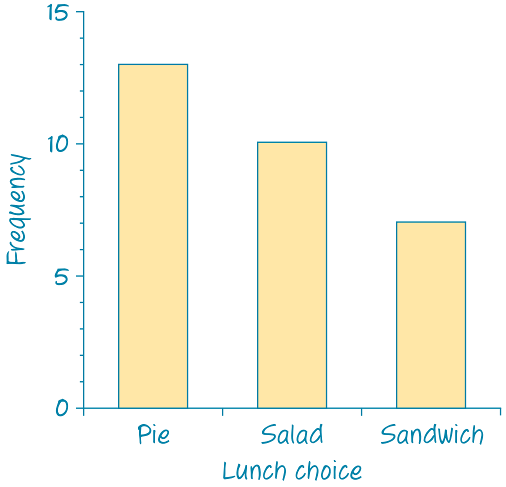
**b** a percentage bar chart.

### Solution

**a** **1** Label the horizontal axis with the variable name, 'Lunch choice'. Mark the scale off into three equal intervals and label them 'Pie', 'Salad' and 'Sandwich'.

**2** Label the vertical axis 'Frequency'. Insert a scale allowing for the maximum frequency of 13. Up to 15 would be appropriate. Mark the scale in 5s.

**3** For each interval draw in a bar as shown. Make the width of each bar less than the width of the category intervals to show that the categories are quite separate. The height of each bar is equal to the frequency.



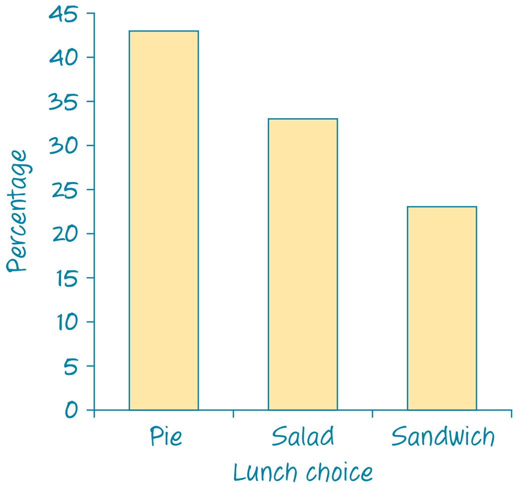
Note: For nominal variables it is common, but not necessary, to list categories in decreasing order by frequency. This makes later interpretation easier.

ISBN 978-1-107-56755-9

Cambridge University Press

**b** To construct a percentage bar chart of the lunch choice data, follow the same procedure as above but label the vertical axis 'Percentage'. Insert a scale allowing for a maximum percentage frequency up to 45%. Mark the vertical scale in intervals of 5%.

The height of each bar is equal to the percentage.



## The mode or modal category

One of the features of a data set that is quickly revealed with a bar chart is the **mode** or **modal category**. This is the most frequently occurring category. In a bar chart, this is given by the category with the tallest bar. For the bar chart in Example 2, the modal category is clearly 'pie'. That is, the most frequent or popular lunch choice was a pie.

### When is the mode useful?

The mode is most useful when a single value or category in the frequency table occurs more often (frequently) than the others. Modes are of particular importance in popularity polls, answering questions like 'Which is the most frequently watched TV station between the hours of 6 p.m. and 8 p.m.?' or 'When is a supermarket in peak demand?'

## Exercise 2B

Skillsheet

### Constructing frequency tables

**1** The *sex* of 15 people in a bus is as shown (F = female, M = male):

F M M M F M F F M M M F M M M

Example 1

**a** Is the variable *sex* nominal or ordinal?

**b** Construct a frequency table for the data including frequencies and percentages frequency.

**2** The UK *shoe size* of 20 eighteen-year-old males are as shown:

|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| 8 | 9 | 9 | 10 | 8 | 8 | 7 | 9 | 8 | 9 |
| 10 | 12 | 8 | 10 | 7 | 8 | 8 | 7 | 11 | 11 |

**a** Is the variable *shoe size* nominal or ordinal?

**b** Construct a frequency table for the data including frequencies and percentages.

ISBN 978-1-107-56755-9

## Analysing frequency tables and constructing bar charts

Example 2

**3** The table below shows the frequency distribution of the favourite type of fast food (*food type*) of a group of students.

| | Frequency | |
|---|---|---|
| Food type | Number | % |
| Hamburgers | 23 | 33.3 |
| Chicken | 7 | 10.1 |
| Fish and chips | 6 | |
| Chinese | 7 | 10.1 |
| Pizza | 18 | |
| Other | 8 | 11.6 |
| Total | | 99.9 |

**a** Complete the table.

**b** Is the variable *food type* nominal or ordinal?

**c** How many students preferred Chinese food?

**d** What percentage of students chose chicken as their favourite fast food?

**e** What was the favourite type of fast food for these students?

**f** Construct a bar chart of the frequencies (number).

**4** The following responses were received to a question regarding the return of capital punishment.

| | Frequency | |
|---|---|---|
| Capital punishment | Number | % |
| Strongly agree | 21 | 8.2 |
| Agree | 11 | 4.3 |
| Don't know | 42 | |
| Disagree | | |
| Strongly disagree | 129 | 50.4 |
| Total | 256 | 100.0 |

**a** Complete the table.

**b** Is the data used to generate this table nominal or ordinal?

**c** How many people said 'Strongly agree'?

**d** What percentage of people said 'Strongly disagree'?

**e** What was the most frequent response?

**f** Construct a frequencies bar chart.

**5** A bookseller noted the types of books purchased during a particular day, with the following results.

| | Frequency | |
|---|---|---|
| Type of book | Number | % |
| Children | 53 | 22.8 |
| Fiction | 89 | |
| Cooking | 42 | 18.1 |
| Travel | 15 | |
| Other | 33 | 14.2 |
| Total | 232 | |

**a** Complete the table.

**b** Is the variable *type of book* nominal or ordinal?

**c** How many books purchased were classified as 'Fiction'?

**d** What percentage of books were classified as 'Children'?

**e** How many books were purchased in total?

**f** Construct a bar chart of the percentage frequencies (%).

**6** A survey of secondary school students' preferred ways of spending their leisure time at home gave the following results.

**a** How many students were surveyed?

**b** Is the variable *leisure activity* nominal or ordinal?

**c** What percentage of students said that they preferred to spend their leisure time phoning a friend?

**d** What was the most popular way of spending their leisure time for these students?

**e** Construct a bar chart of the percentage frequencies (%).

| | Frequency | |
|---|---|---|
| Leisure activity | Number | % |
| Watch TV | 84 | 42 |
| Read | 26 | 13 |
| Listen to music | 46 | 23 |
| Watch a movie | 24 | 12 |
| Phone friends | 8 | 4 |
| Other | 12 | 6 |
| Total | 200 | 100 |

## 2C Interpreting and describing frequency tables and bar charts

As part of this subject, you will be expected to complete a statistical investigation. Under these circumstances, constructing a frequency table or a bar chart is not an end in itself. It is merely a means to an end. The end is being able to understand something about the variables you are investigating that you didn't know before.

To complete the investigation, you will need to communicate this finding to others. To do this, you will need to know how to describe and interpret any patterns you observe in the context of your data investigation in a written report that is both systematic and concise. The purpose of this section is to help you develop such skills.

### Some guidelines for describing the distribution of a categorical variable and communicating your findings

- Briefly summarise the context in which the data were collected including the number of people (or things) involved in the study.
- If there is a clear modal category, make sure that it is mentioned.
- Include relevant counts or percentages in the report.
- If there are a lot of categories, it is not necessary to mention every category.
- Either counts or percentages can be used to describe the distribution.

These guidelines are illustrated in the following examples.

ISBN 978-1-107-56755-9

Cambridge University Press

### Example 3 Using a frequency table to describe the outcome of an investigation involving a categorical variable

A group of 30 children were offered a choice of a sandwich, a salad or a pie for lunch and their responses collected and summarised in the frequency table opposite.

| Lunch choice | Frequency |
|---|---|
| Sandwich | 7 |
| Salad | 10 |
| Pie | 13 |
| Total | 30 |

Use the frequency table to report on the relative popularity of the three lunch choices quoting appropriate frequencies to support your conclusions.

**Solution**

Report

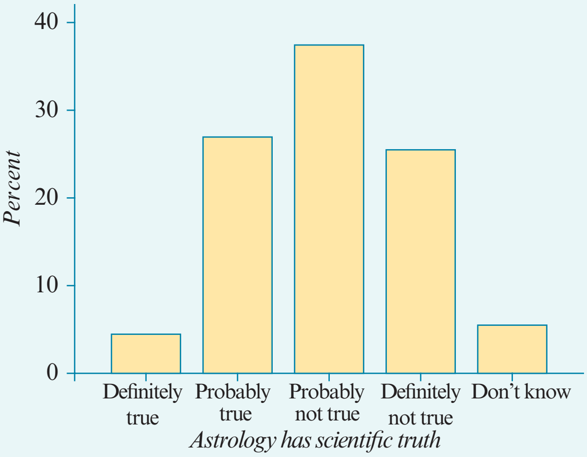
A group of 30 children were offered a choice of a sandwich, a salad or a pie for lunch. The most popular lunch choice was pie, chosen by 13 of the children. Ten of the children chose a salad. The least popular option was sandwich, chosen by only 7 of the children.

### Example 4 Using a frequency table and a percentage bar chart to describe the outcome of an investigation involving a categorical variable

A sample of 200 people were asked to comment on the statement 'Astrology has scientific truth' by selecting one of the options 'definitely true', 'probably true', 'probably not true', definitely not true' or 'don't know'.

The data are summarised in the following frequency table and bar chart. Note that the categories in the frequency table can be ordered in a definite order because the data are ordinal.

| Astrology has scientific truth | Frequency | |
|---|---|---|
| | Number | % |
| Definitely true | 9 | 4.5 |
| Probably true | 54 | 27.0 |
| Probably not true | 75 | 37.5 |
| Definitely not true | 51 | 25.5 |
| Don't know | 11 | 5.5 |
| Total | 200 | 100.0 |



Write a report summarising the findings of this investigation quoting appropriate percentages to support your conclusion.

### Solution

Report

Two hundred people were asked to respond to the statement 'Astrology has scientific truth'.

The majority of respondents did not agree, with 37.5% responding that they believed that this statement was probably not true, and another 25.5% declaring that the statement was definitely not true. Over one quarter (27%) of the respondents thought that the statement was probably true, while only 4.5% of subjects thought that the statement was definitely true.

## Exercise 2C

### Interpreting and describing frequency tables and bar charts

Example 3

**1** A group of 69 students were asked to nominate their preferred type of fast food. The results are summarised in the percentage frequency table opposite. Use the information in the table to complete the report below by filling in the blanks.

| Fast food type | % |
|---|---|
| Hamburgers | 33.3 |
| Chicken | 10.1 |
| Fish and chips | 8.7 |
| Chinese | 10.1 |
| Pizza | 26.1 |
| Other | 11.6 |
| Total | 99.9 |

Report

A group of □ students were asked their favourite type of fast food. The most popular response was □ (33.3%), followed by pizza (□). The rest of the group were almost evenly split between chicken, fish and chips, Chinese and other, all around 10%.

**2** Two hundred and fifty-six people were asked whether they agreed that there should be a return to capital punishment in their state. Their responses are summarised in the table opposite. Use the information in the table to complete the report below by filling in the blanks.

| Capital punishment | % |
|---|---|
| Strongly agree | 8.2 |
| Agree | 4.3 |
| Don't know | 16.4 |
| Disagree | 20.7 |
| Strongly disagree | 50.4 |
| Total | 100.0 |

Report

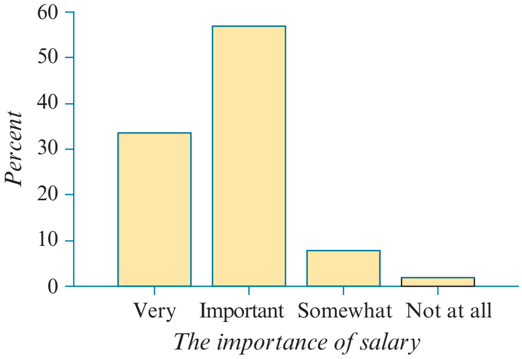
A group of 256 people were asked whether they agreed that there should be a return to capital punishment in their state. The majority of these people □ (50.4%), followed by □ who disagreed. Levels of support for return to capital punishment were quite low, with only 4.3% agreeing and 8.2% strongly agreeing. The remaining □ said that they didn't know.

**3** A group of 200 students were asked how they prefer to spend their leisure time. The results are summarised in the frequency table below.
Use the information in the table to write a brief report of the results of this investigation.

| Leisure activity | % |
|---|---|
| Internet and digital games | 42 |
| Read | 13 |
| Listen to music | 23 |
| Watch TV or go to movies | 12 |
| Phone friends | 4 |
| Other | 6 |
| Total | 100 |

Example 4 **4** A group of 579 employees from a large company were asked to rate the importance of salary in determining how they felt about their job. Their responses are shown in the following frequency table and bar chart.

| Importance of salary | % |
|---|---|
| Very important | 33.5 |
| Important | 56.8 |
| Somewhat important | 7.8 |
| Not at all important | 1.9 |
| *Total* | 100.0 |



Write a report describing how these employees rated the importance of salary in determining how they felt about their job.

## 2D Displaying and describing numerical data

Frequency tables can also be used to organise numerical data. For discrete numerical data, the process is the same as that for categorical data, as shown in the following example.

### ▶ Discrete data

**Example 5** **Constructing a frequency table for discrete numerical data**

The number of brothers and sisters (siblings) reported by each of the 30 students in year 11 are as follows:

2 3 4 0 3 2 3 0 4 1 0 0 1 2 3
0 2 1 1 4 5 3 2 5 6 1 1 1 0 2

Construct a frequency table for these data.

**Solution**

**1** Find the maximum and the minimum values in the data set. Here the minimum is 0 and the maximum is 6.

**2** Construct a table as shown, including all the values between the minimum and the maximum.

**3** Count the number of 0s, 1s, 2s, etc. in the data set. For example, there are seven 1s. Record these values in the number column.

**4** Add the frequencies to find the total.

**5** Convert the frequencies to percentages, and record in the per cent (%) column.

For example, percentage of 1s equals $\frac{7}{30} \times 100 = 23.3\%$.

**6** Total the percentages and record.

| Number of siblings | Frequency | |
|---|---|---|
| | Number | % |
| 0 | 6 | 20.0 |
| 1 | 7 | 23.3 |
| 2 | 6 | 20.0 |
| 3 | 5 | 16.7 |
| 4 | 3 | 10.0 |
| 5 | 2 | 6.7 |
| 6 | 1 | 3.3 |
| Total | 30 | 100.0 |

### ▶ Grouping data

Some variables can only take on a limited range of values, for example, the variable *number of children in a family*. For these variables, it makes sense to list each of these values individually when forming a frequency distribution.

In other cases, the variable can take on a large range of values: for example, the variable *age* might take values from 0 to 100 or even more. Listing all possible ages would be tedious and would produce a large and unwieldy table. To solve this problem we **group the data** into a small number of convenient intervals.

ISBN 978-1-107-56755-9

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.
- 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.

## Grouped discrete data

### Example 6 Constructing a frequency table for a discrete numerical variable

A group of 20 people were asked to record how many cups of coffee they drank in a particular week, with the following results:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0 | 9 | 10 | 23 | 25 | 0 | 0 | 34 | 32 |
| 5 | 0 | 17 | 14 | 3 | 6 | 0 | 33 | 23 | 0 |

Construct a table of these data showing both frequency (count) and percentage frequency.

**Solution**

**1** The minimum number of cups of coffee drunk is 0 and the maximum is 34. Intervals beginning at 0 and ending at 34 would ensure that all the data are included. Interval width of 5 will mean that there are 7 intervals. Note that the endpoints are within the interval, so that the interval 0 – 4 includes 5 values: 0, 1, 2, 3, 4.

**2** Set up the table as shown.

**3** Count the number of data values in each interval to complete the number column.

| | Frequency | |
|---|---|---|
| Cups of coffee | Number | % |
| 0–4 | 8 | 40 |
| 5–9 | 3 | 15 |
| 10–14 | 2 | 10 |
| 15–19 | 1 | 5 |
| 20–24 | 2 | 10 |
| 25–29 | 1 | 5 |
| 30–34 | 3 | 15 |
| Total | 20 | 100 |

**4** Convert the frequencies into percentages and record in the per cent (%) column. For example, for the interval 5–9: % frequency $= \frac{3}{20} \times 100 = 15\%$

**5** Total the percentages and record.

ISBN 978-1-107-56755-9

Cambridge University Press

## Grouped continuous data

### Example 7 Constructing a frequency table for a continuous numerical variable

The following are the heights of the 41 players in a basketball club, in centimetres.

| | | | | |
|---|---|---|---|---|
| 178.1 | 185.6 | 173.3 | 193.4 | 183.1 |
| 184.6 | 202.4 | 170.9 | 183.3 | 180.3 |
| 185.8 | 189.1 | 178.6 | 194.7 | 185.3 |
| 191.1 | 189.7 | 191.1 | 180.4 | 180.0 |
| 193.8 | 196.3 | 189.6 | 183.9 | 177.7 |
| 178.9 | 193.0 | 188.3 | 189.5 | 182.0 |
| 183.6 | 184.5 | 188.7 | 192.4 | 203.7 |
| 180.1 | 170.5 | 179.3 | 184.1 | 183.8 |
| 174.7 | | | | |

Construct a frequency table of these data.

**Solution**

1 Find the minimum and maximum heights, which are 170.5 cm and 203.7 cm. A minimum value of 170 and a maximum of 204.9 will ensure that all the data are included.

2 Interval width of 5 cm will mean that there are 7 intervals from 170 to 204.9, which is within the guidelines of 5–15 intervals.

3 Set up the table as shown. All values of the variable that are from 170 to 174.9 have been included in the first interval. The second interval includes values from 175 to 179.9, and so on for the rest of the table.

| Heights | Frequency | |
|---|---|---|
| | Number | % |
| 170–174.9 | 4 | 9.8 |
| 175–179.9 | 5 | 12.2 |
| 180–184.9 | 13 | 31.7 |
| 185–189.9 | 9 | 22.0 |
| 190–194.9 | 7 | 17.1 |
| 195–199.9 | 1 | 2.4 |
| 200–204.9 | 2 | 4.9 |
| Total | 41 | 100.1 |

4 The number of data values in each interval is then counted to complete the number column of the table.

5 Convert the frequencies into percentages and record in the per cent (%) column. For example, for the interval 175.0–179.9: % frequency $= \frac{5}{41} \times 100 = 12.2\%$

6 Total the percentages and record.

The interval that has the highest frequency is called the **modal interval**. Here the modal interval is 180.0–184.9, as 13 players (31.7%) have heights that fall into this interval.

ISBN 978-1-107-56755-9

Cambridge University Press

## Histograms

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.

### Histograms

In a histogram:

- frequency (number or percentage) is shown on the vertical axis
- the values of the variable being displayed are plotted on the horizontal axis
- each column corresponds to a data value, or a data interval if the data is grouped; alternatively, for ungrouped discrete data, the actual data value is located at the middle of the column
- the height of the column gives the frequency (number or percentage).

### Example 8 Constructing a histogram for ungrouped discrete data

Construct a histogram for the data in the frequency table.

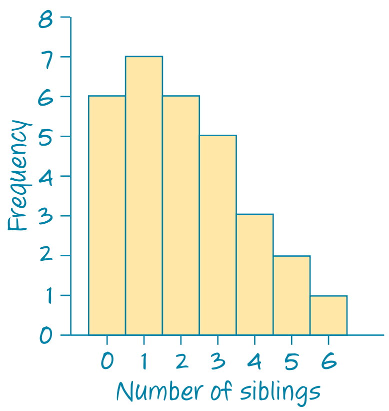
| Siblings | Frequency |
|---|---|
| 0 | 6 |
| 1 | 7 |
| 2 | 6 |
| 3 | 5 |
| 4 | 3 |
| 5 | 2 |
| 6 | 1 |
| Total | 30 |

#### Solution

1 Label the horizontal axis with the variable name 'Number of siblings'. Mark in the scale in units, so that it includes all possible values.

2 Label the vertical axis 'Frequency'. Insert a scale allowing for the maximum frequency of 7. Up to 8 would be appropriate. Mark the scale in units.

3 For each value for the variable draw in a column. The data is discrete, so make the width of each column 1, starting and ending halfway between data values. For example, the column representing 2 siblings starts at 1.5 and ends at 2.5. The height of each column is equal to the frequency.



## Example 9 Constructing a histogram for continuous data

Construct a histogram for the data in the frequency table.

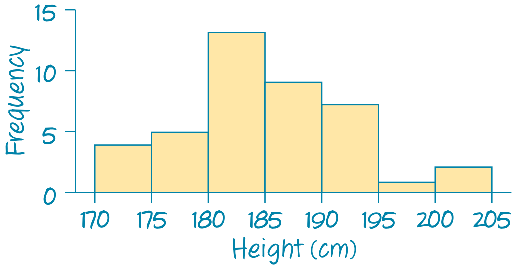
| Height (cm) | Frequency |
|---|---|
| 170.0–174.9 | 4 |
| 175.0–179.9 | 5 |
| 180.0–184.9 | 13 |
| 185.0–189.9 | 9 |
| 190.0–194.9 | 7 |
| 195.0–199.9 | 1 |
| 200.0–204.9 | 2 |
| Total | 41 |

### Solution

**1** Label the horizontal axis with the variable name ‘Height’. Mark in the scale using the beginning of each interval as the scale points; that is, 170, 175, . . .

**2** Label the vertical axis ‘Frequency’. Insert a scale allowing for the maximum frequency of 13. Up to 15 would be appropriate. Mark the scale in units.

**3** For each interval draw in a column. Each column starts at the beginning of the interval and finishes at the beginning of the next interval. Make the height of each column equal to the frequency.



## Constructing a histogram using a CAS calculator

It is relatively quick to construct a histogram from a frequency table. However, if you only have the 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 us.

ISBN 978-1-107-56755-9

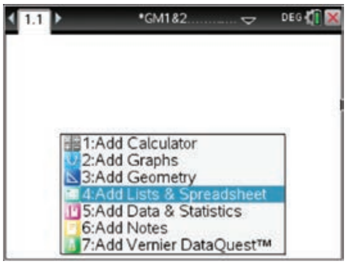
## How to construct a histogram using the TI-Nspire CAS

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

### Steps

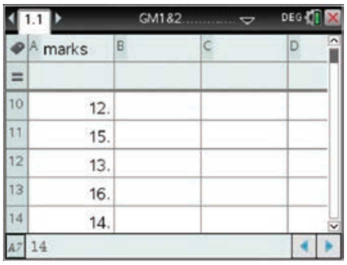
1 Start a new document: Press ⌂on and select **New Document** (or use ctrl + N). If prompted to save an existing document, move cursor to **No** and press enter.



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

a Move the cursor to the name space of column A (or any other column) and type in *marks* as the list name. Press enter.

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

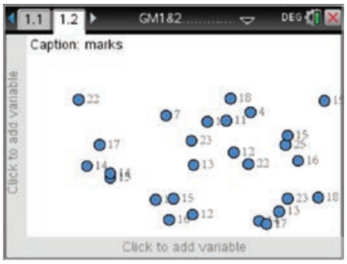


| | A marks | B | C | D |
|---|---|---|---|---|
| = | | | | |
| 10 | 12. | | | |
| 11 | 15. | | | |
| 12 | 13. | | | |
| 13 | 16. | | | |
| 14 | 14. | | | |

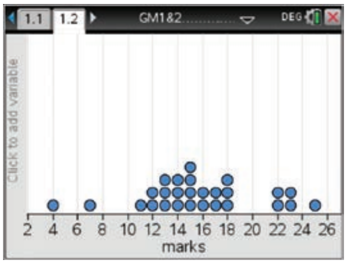
A7 14

3 Statistical graphing is done through the **Data & Statistics** application.
Press ctrl + I and select **Add Data & Statistics** (or press ⌂on, arrow to , and press enter).

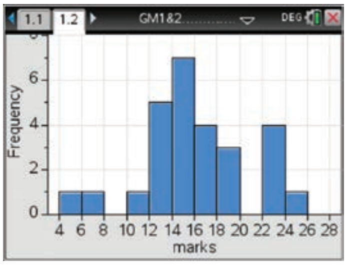
Note: A random display of dots will appear – this is to indicate that data are available for plotting. It is not a statistical plot.



a Press tab to show the list of variables that are available. Select the variable **marks**. Press enter to paste the variable **marks** to that axis.



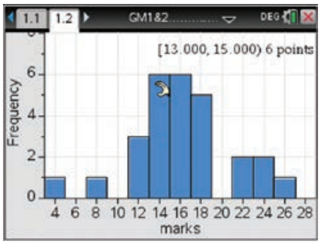
b A dot plot is displayed as the default plot. To change the plot to a histogram, press menu>**Plot Type>Histogram** and then press enter or 'click' (press ).
Your screen should now look like that shown opposite. This histogram has a column (or bin) width of 2 and a starting point of 3.



ISBN 978-1-107-56755-9

Cambridge University Press

**4** Data analysis



**a** Move cursor onto 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. To deselect any previously selected columns move the cursor to the open area and press ☝.

Hint: If you accidentally move a column or data point, press `ctrl` + `esc` to undo the move.

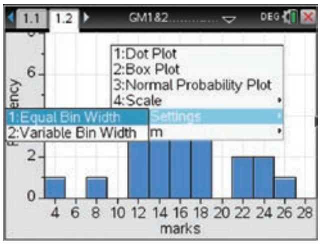
**5** Change the histogram column (bin) width to 4 and the starting point to 2.

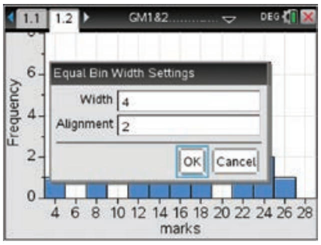
**a** Press `ctrl` + `menu` to get the contextual menu as shown (below left).

Hint: Pressing `ctrl` + `menu` with the cursor on the histogram gives you access to a contextual menu that enables you to do things that relate only to histograms.

**b** Select **Bin Settings**.

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

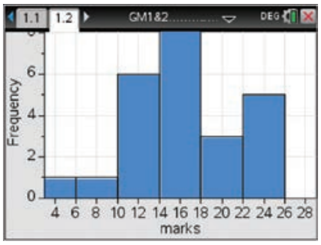


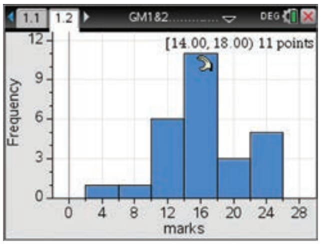

Equal Bin Width Settings

Width 4

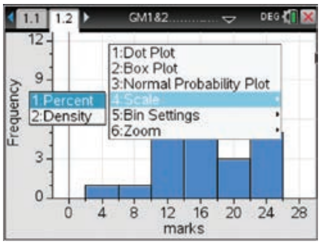
Alignment 2

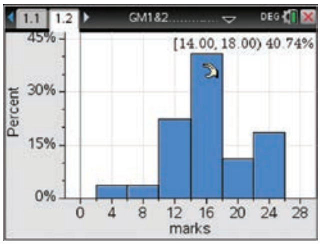
**d** A new histogram is displayed with column width of 4 and a starting point of 2 but it no longer fits the viewing window (below left). To solve this problem press `ctrl` + `menu` **>Zoom>Zoom-Data** and `enter` to obtain the histogram as shown below right.





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





## How to construct a histogram using the ClassPad

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

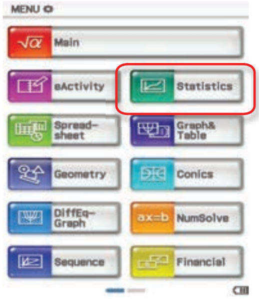
16 11 4 25 15 7 14 13 14 12 15 13 16 14
15 12 18 22 17 18 23 15 13 17 18 22 23

### Steps

**1** From the application menu screen, locate the **Statistics** application.

Tap Statistics to open.

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



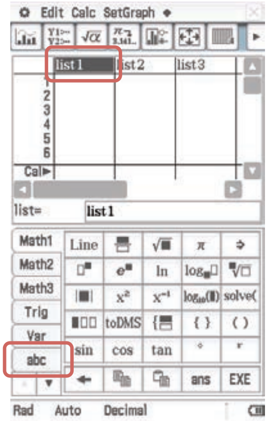
**2** Enter the data into a list named **marks**.

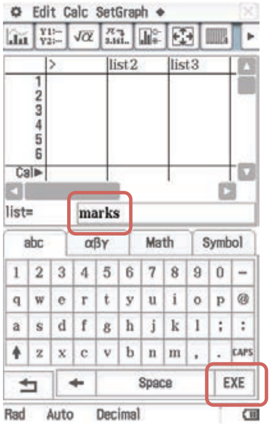
**a** Highlight the heading of the first list by tapping.

**b** Press Keyboard and tap abc.

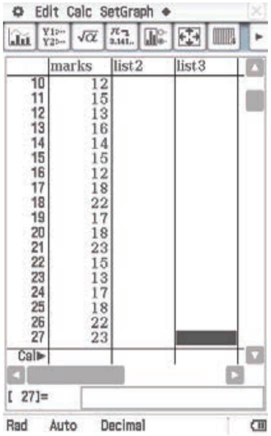
**c** Type **marks** and press EXE.

**d** Starting in row 1, type in each data value. Press EXE or ▼ to move down the list.





Your screen should be like the one shown at right.



| | marks | list2 | list3 |
|---|---|---|---|
| 10 | 12 | | |
| 11 | 15 | | |
| 12 | 13 | | |
| 13 | 16 | | |
| 14 | 14 | | |
| 15 | 15 | | |
| 16 | 12 | | |
| 17 | 18 | | |
| 18 | 22 | | |
| 19 | 17 | | |
| 20 | 18 | | |
| 21 | 23 | | |
| 22 | 15 | | |
| 23 | 13 | | |
| 24 | 17 | | |
| 25 | 18 | | |
| 26 | 22 | | |
| 27 | 23 | | |

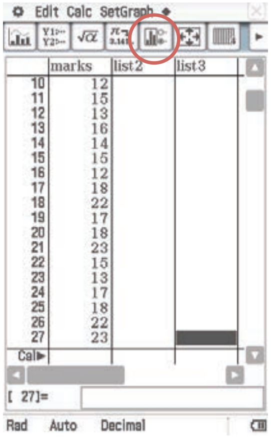
**3** To plot a statistical graph:

**a** Tap at the top of the screen. This opens the **Set StatGraphs** dialog box.

**b** Complete the dialog box. For:

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

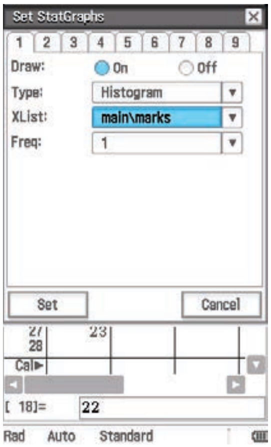
**c** Tap Set to confirm your selections.



| | marks | list2 | list3 |
|---|---|---|---|
| 10 | 12 | | |
| 11 | 15 | | |
| 12 | 13 | | |
| 13 | 16 | | |
| 14 | 14 | | |
| 15 | 15 | | |
| 16 | 12 | | |
| 17 | 18 | | |
| 18 | 22 | | |
| 19 | 17 | | |
| 20 | 18 | | |
| 21 | 23 | | |
| 22 | 15 | | |
| 23 | 13 | | |
| 24 | 17 | | |
| 25 | 18 | | |
| 26 | 22 | | |
| 27 | 23 | | |

[ 27]=

Rad Auto Decimal



Set StatGraphs

Draw: ● On ○ Off

Type: Histogram

XList: main\marks

Freq: 1

Set Cancel

[ 18]= 22

Rad Auto Standard

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

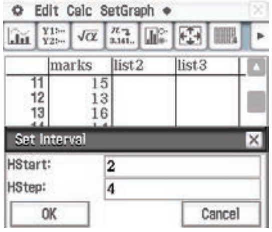
**4** To plot the graph:

**a** Tap in the toolbar.

**b** Complete the **Set Interval** dialog box as given below. For:

- **HStart:** type in **2**
- **HStep:** type in **4**.

**c** Tap **OK**.



| | marks | list2 | list3 |
|---|---|---|---|
| 11 | 15 | | |
| 12 | 13 | | |
| 13 | 16 | | |

Set Interval

HStart: 2

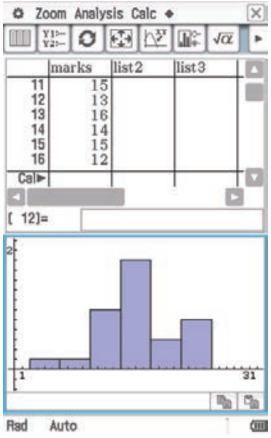
HStep: 4

OK Cancel

**5** The screen is split in two. Tapping Resize from the icon panel will allow the graph to fill the entire screen.
Tap Resize to return to half-screen size.

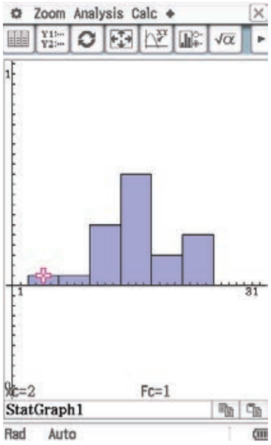
**6** Tapping places a marker on the first column of the histogram and tells us that:

- the first interval begins at 2 (**$x_c$ = 2**)
- for this interval, the frequency is 1 (**$F_c$ = 1**).



| | marks | list2 | list3 |
|---|---|---|---|
| 11 | 15 | | |
| 12 | 13 | | |
| 13 | 16 | | |
| 14 | 14 | | |
| 15 | 15 | | |
| 16 | 12 | | |

[ 12]=



xc=2 Fc=1

StatGraph1

Rad Auto

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

# Exercise 2D

## Constructing frequency tables for numerical data

Example 5 **1** The number of magazines purchased in a month by 15 different people was as follows:

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

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

Example 6 **2** The amount of money carried by 20 students is as follows:

$4.55 $1.45 $16.70 $0.60 $5.00 $12.30 $3.45 $23.60 $6.90 $4.35
$0.35 $2.90 $1.70 $3.50 $8.30 $3.50 $2.20 $4.30 $0.00 $11.50

Construct a frequency table for the data, including both the number and percentage in each category. Use intervals of $5, starting at $0.

## Analysing frequency tables and constructing histograms

Example 7 **3** A group of 28 students were asked to draw a line that they estimated to be the same length as a 30 cm ruler. The results are shown in the frequency table below.

| | Frequency | |
|---|---|---|
| Length of line (cm) | Number | % |
| 28.0–28.9 | 1 | 3.6 |
| 29.0–29.9 | 2 | 7.1 |
| 30.0–30.9 | 8 | 28.6 |
| 31.0–31.9 | 9 | 32.1 |
| 32.0–32.9 | 7 | 25.0 |
| 33.0–33.9 | 1 | 3.6 |
| Total | 28 | 100.0 |

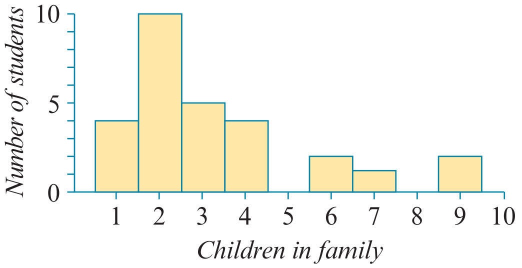
**a** How many students drew a line with a length:
- **i** from 29.0 to 29.9 cm?
- **ii** of less than 30 cm?
- **iii** of 32 cm or more?

**b** What percentage of students drew a line with a length:
- **i** from 31.0 to 31.9 cm?
- **ii** of less than 31 cm?
- **iii** of 30 cm or more?

**c** Use the table to construct a histogram using the counts.

## Interpreting histograms

**4** The number of children in the family for each student in a class is shown in the histogram.



**a** How many students are the only child in a family?

**b** What is the most common number of children in a family?

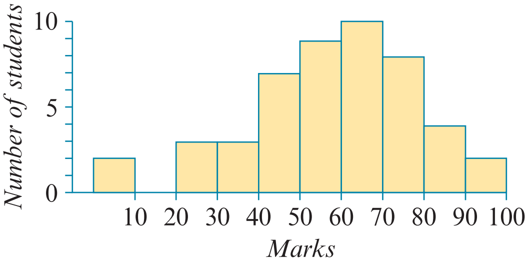
**c** How many students come from families with 6 or more children?

**d** How many students are there in the class?

ISBN 978-1-107-56755-9

Cambridge University Press

**5** The following histogram gives the scores on a general knowledge quiz for a class of year 11 students.



**a** How many students scored from 10 to 19 marks?

**b** How many students attempted the quiz?

**c** What is the modal interval?

**d** If a mark of 50 or more is designated as a pass, how many students passed the quiz?

## Constructing histograms using a CAS calculator and their analysis

Example 8

**6** A student purchased 21 new textbooks from a schoolbook supplier with the following prices (in dollars):

41.65 34.95 32.80 27.95 32.50 53.99 63.99 17.80 13.50 18.99 42.98
38.50 59.95 13.20 18.90 57.15 24.55 21.95 77.60 65.99 14.50

**a** Use a CAS calculator to construct a histogram with column width 10 and starting point 10. Name the variable *price*.

**b** For this histogram:

**i** what is the range of the third interval?

**ii** what is the 'frequency' for the third interval?

**iii** what is the modal interval?

**7** The maximum temperatures for several capital cities around the world on a particular day, in degrees Celsius, were:

17 26 36 32 17 12 32 2 16 15 18 25
30 23 33 33 17 23 28 36 45 17 19 37

**a** Use a CAS calculator to construct a histogram with column width 2 and starting point 0. Name the variable *maxtemp*.

**b** For this histogram:

**i** what is the starting point of the second column?

**ii** what is the 'frequency' for this interval?

**c** Use the window menu to redraw the histogram with a column width of 5 and a starting point of 0.

**d** For this histogram:

**i** how many cities had maximum temperatures from 20°C to 25°C?

**ii** what is the modal interval?

# 2E Characteristics of distributions of numerical data: shape, location and spread

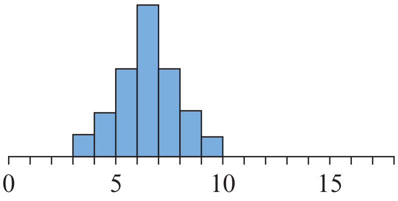
Distributions of numerical data are characterised by their shape and special features such as centre and spread.

## Shape of a distribution

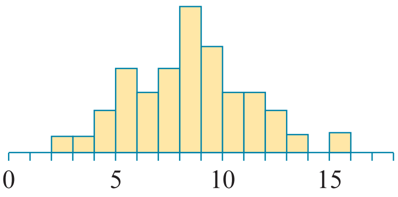
### Symmetry and skew

A distribution is said to be **symmetric** if it forms a mirror image of itself when folded in the 'middle' along a vertical axis. Otherwise, the distribution is **skewed**.

Histogram A is symmetric, while Histogram B shows a distribution that is approximately symmetric.



Histogram A



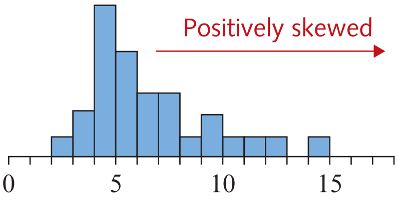
Histogram B

### Positive and negative skew

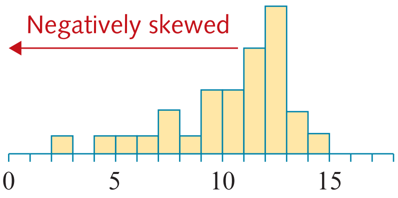
A histogram may be positively or negatively skewed.

- It is **positively skewed** if it has a short tail to the left and a long tail pointing to the right (because of the many values towards the positive end of the distribution).
- It is **negatively skewed** if it has a short tail to the right and a long tail pointing to the left (because of the many values towards the negative end of the distribution).

Histogram C is an example of a positively skewed distribution, and Histogram D is an example of a negatively skewed distribution.



Histogram C



Histogram D

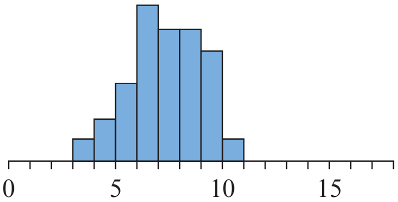
Knowing whether a distribution is skewed or symmetric is important, as this gives considerable information concerning the choice of appropriate summary statistics, as will be seen in the next section.

## ▶ Location and spread

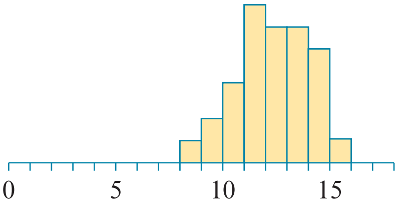
### Comparing location

Two distributions are said to differ in **location** if the values of the data in one distribution are generally larger than the values of the data in the other distribution.

Consider, for example, the following histograms, shown on the same scale. Histogram F is identical in shape and width to Histogram E but moved horizontally several units to the right, indicating that these distributions differ in location.



Histogram E

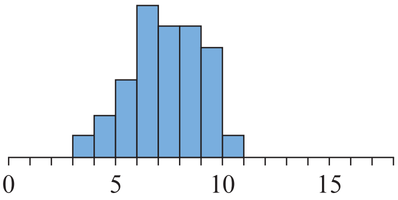


Histogram F

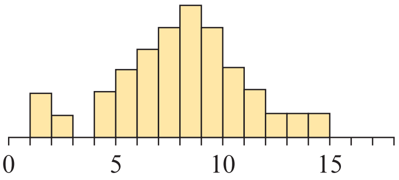
### Comparing spread

Two distributions are said to differ in **spread** if the values of the data in one distribution tend to be more variable (spread out) than the values of the data in the other distribution.

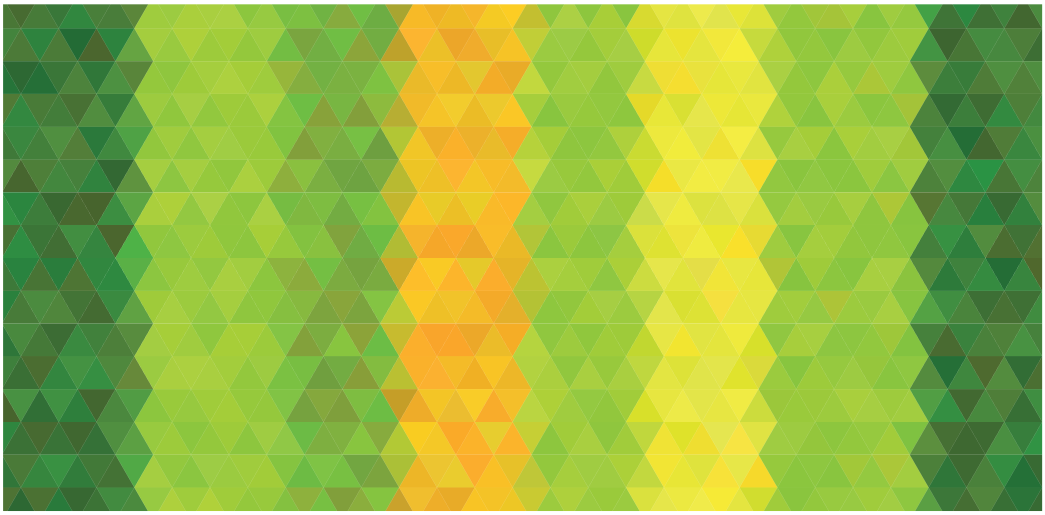
Histograms G and H illustrate the difference in spread. While both are centred at about the same place, Histogram H is more spread out.



Histogram G



Histogram H



ISBN 978-1-107-56755-9

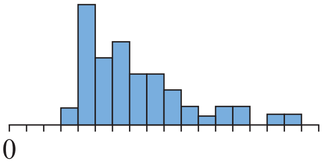
Cambridge University Press

## Exercise 2E

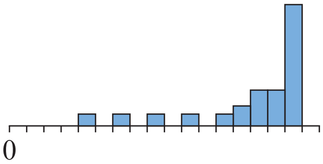
### Describing shape using histograms

1 Describe the shape of each of the following histograms.

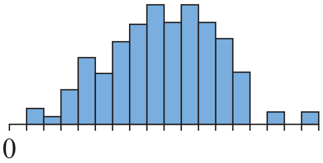
a



b

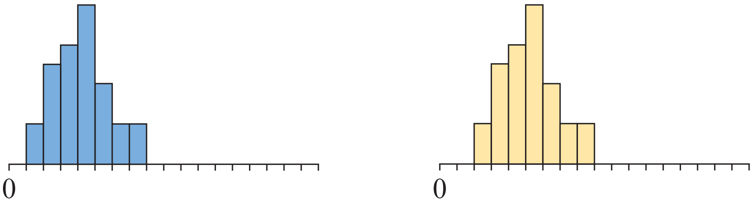


c

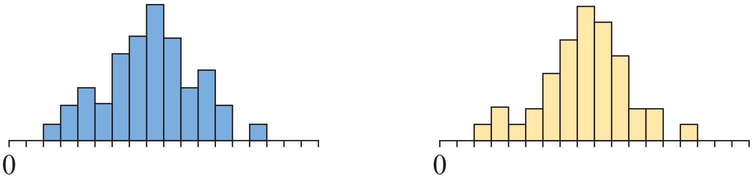


2 Do the following pairs of distributions differ in spread, location, both or neither? Assume that each pair of histograms is drawn on the same scale.

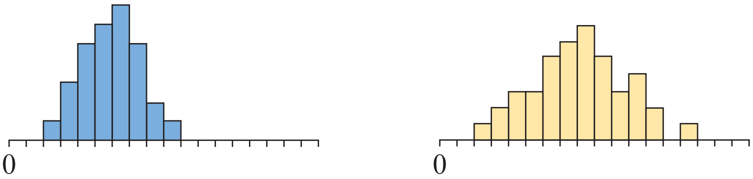
a



b



c



## 2F Dot plots and stem-and-leaf plots

### ▶ Dot plots

The simplest display of numerical data is a **dot plot**.

**Dot plot**

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.

Dot plots are a great way to display fairly small data sets where the data takes a limited number of values.

ISBN 978-1-107-56755-9

Cambridge University Press

### Example 10 Constructing a dot plot

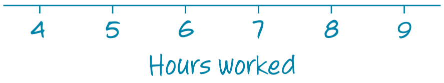
The number of hours worked by each of 10 students in their part-time jobs is as follows:

6 9 5 8 6 4 6 7 6 5

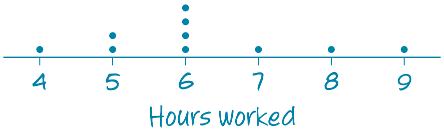
Construct a dot plot of these data.

**Solution**

**1** Draw in a number line, scaled to include all data values. Label the line with the variable being displayed.



**2** Plot each data value by marking in a dot above the corresponding value on the number as shown.



## Stem-and-leaf plots

The **stem-and-leaf plot** or **stem plot** is another plot used for small data sets.

### Example 11 Constructing a stem plot

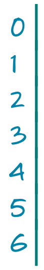
The following is a set of marks obtained by a group of students on a test:

15 2 24 30 25 19 24 33 41 60 42 35 35
28 28 19 19 28 25 20 36 38 43 45 39

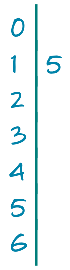
Display the data in the form of an ordered stem-and-leaf plot.

**Solution**

**1** The data set has values in the units, tens, twenties, thirties, forties, fifties and sixties. Thus, appropriate stems are 0, 1, 2, 3, 4, 5 and 6. Write these down in ascending order, followed by a vertical line.



**2** Now attach the leaves. The first data value is 15. The stem is 1 and the leaf is 5. Opposite the 1 in the stem, write the number 5, as shown.



The second data value is 2. The stem is 0 and the leaf is 2. Opposite the 0 in the stem, write the number 2, as shown.

```
0 | 2
1 | 5
2 |
3 |
4 |
5 |
6 |
```

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

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

**3** Ordering the leaves in increasing value as they move away from the stem gives the *ordered* stem plot, as shown. Write the name of the variable being displayed (*Marks*) at the top of the plot and add a key (1|5 means 15 marks).

```
Marks        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
```

It can be seen from this plot that the distribution is approximately symmetric, with one test score, 60, which seems to stand out from the rest. When a value sits away from the main body of the data, it is called an **outlier**.

## ▶ Choosing between plots

We now have three different plots that can be used to display numerical data: the histogram, the dot plot and the stem plot. They all allow us to make judgements concerning the important features of the distribution of the data, so how would we decide which one to use?

While there are no hard and fast rules, the following guidelines are often used.

| Plot | Used best when | How usually constructed |
|---|---|---|
| Dot plot | small data sets (say $n < 30$)<br>discrete data | by hand or with technology when constructing histograms as well |
| Stem plot | small data sets (say $n < 50$) | by hand |
| Histogram | large data sets (say $n > 30$) | with technology |

## Exercise 2F

### Constructing and analysing dot plots

Example 10

**1** The number of children in each of 15 families is as follows:

0 7 2 2 2 4 1 3 3 2 2 2 0 0 1

**a** Construct a dot plot of the number of children.

**b** What is the mode of this distribution?

**2** A group of 20 people were asked how many times in the last week they had shopped at a particular supermarket. Their responses were as follows:

0 1 1 0 0 6 0 1 2 2
3 4 0 0 1 1 2 3 2 0

**a** Construct a dot plot of this data.

**b** How many people did not shop at the supermarket in the last week?

**3** The number of goals scored in an AFL game by each player on one team is as follows:

0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 1 1 1 2 2 3 6

**a** Construct a dot plot of the number of goals scored.

**b** What is the mode of this distribution?

**c** What is the shape of the distribution of goals scored?

**4** In a study of the service offered at her café, Amanda counted the number of people waiting in the queue every 5 minutes from 12 noon until 1 p.m.:

| *Time* | 12:00 | 12:05 | 12:10 | 12:15 | 12:20 | 12:25 | 12:30 | 12:35 | 12:40 | 12:45 | 12:50 | 12:55 | 1:00 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Number* | 0 | 2 | 4 | 4 | 7 | 8 | 6 | 5 | 0 | 1 | 2 | 1 | 1 |

**a** Construct a dot plot of the number of people waiting in the queue.

**b** When does the peak demand at the café seem to be?

ISBN 978-1-107-56755-9

Cambridge University Press

## Constructing and analysing stem plots

Example 11

**5** The marks obtained by a group of students on an English examination are as follows:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 92 | 65 | 35 | 89 | 79 | 32 | 38 | 46 | 26 | 43 | 83 | 79 |
| 50 | 28 | 84 | 97 | 69 | 39 | 93 | 75 | 58 | 49 | 44 | 59 |
| 78 | 64 | 23 | 17 | 35 | 94 | 83 | 23 | 66 | 46 | 61 | 52 |

**a** Construct a stem plot of the marks.

**b** How many students obtained 50 or more marks?

**c** What was the lowest mark?

**6** 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 43 years old?

Age (years) 1 | 2 represents 12 years

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

**7** An investigator recorded the amount of time for which 24 similar batteries lasted in a toy. Her results (in hours) were:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | 40 | 30 | 24 | 27 | 31 | 21 | 27 | 20 | 30 | 33 | 22 |
| 4 | 26 | 17 | 19 | 46 | 34 | 37 | 28 | 25 | 31 | 41 | 33 |

**a** Make a stem plot of these times.

**b** How many of the batteries lasted for more than 30 hours?

**8** The amount of time (in minutes) that a class of students spent on homework on one particular night were:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 27 | 46 | 63 | 20 | 33 | 15 | 21 | 16 | 14 | 15 |
| 39 | 70 | 19 | 37 | 56 | 20 | 28 | 23 | 0 | 29 | 10 |

**a** Make a stem plot of these times.

**b** How many students spent more than 60 minutes on homework?

**c** What is the shape of the distribution?

**9** The prices of a selection of shoes at a discount outlet are as follows:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $49 | $75 | $68 | $79 | $75 | $39 | $35 | $52 | $149 | $84 |
| $36 | $95 | $28 | $25 | $78 | $45 | $46 | $76 | $82 | |

**a** Construct a stem plot of this data.

**b** What is the shape of the distribution?

ISBN 978-1-107-56755-9

Cambridge University Press

# 2G Summarising data

A statistic is any number computed from data. Certain special statistics are called **summary statistics**, because they numerically summarise important features of the data set. Of course, whenever any set of data is summarised into just one or two numbers, much information is lost. However, if a summary statistic is well chosen, it may reveal important information hidden in the data set.

For a single data distribution, the most commonly used summary statistics are either measures of centre or measures of spread.

## ▶ Measures of centre

### The mean

The most commonly used measure of the centre of a distribution of a numerical variable is the **mean**. The mean is calculated by summing the data values and then dividing by their number. The mean of a set of data is what many people call the 'average'.

### The mean

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

For example, consider the set of data: 1, 5, 2, 4

$$\text{Mean} = \frac{1+5+2+4}{4} = \frac{12}{4} = 3$$

### 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 use a shorthand notation that enables complex statistical formulas to be written out in a compact form.

In this notation we use:

- 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
- $n$ to represent the total number of data values.

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

## Example 12 Calculating the mean

The following data set shows the number of premierships won by each of the current AFL teams, until the end of 2014. Find the mean of the number of premierships won.

| Team | Premierships won |
|---|---|
| Carlton | 16 |
| Essendon | 16 |
| Collingwood | 15 |
| Melbourne | 12 |
| Hawthorn | 12 |
| Brisbane Lions | 11 |
| Richmond | 10 |
| Geelong | 9 |
| Sydney | 5 |
| Kangaroos | 4 |
| West Coast | 3 |
| Adelaide | 2 |
| Port Adelaide | 1 |
| Western Bulldogs | 1 |
| St Kilda | 1 |
| Fremantle | 0 |
| Gold Coast | 0 |
| GWS | 0 |

### Solution

**1** Write down the formula and the value of *n*.

$$\bar{x} = \frac{\sum x}{n} \qquad n = 18$$

**2** Substitute into the formula and evaluate.

$$\bar{x} = \frac{16 + 16 + 15 + ::: + 1 + 1 + 0 + 0 + 0}{18}$$

**3** We do not expect the mean to be a whole number, so give your answer to one decimal place.

$$= \frac{118}{18}$$

$$= 6.6$$

ISBN 978-1-107-56755-9 © Jones et al. 2016 Cambridge University Press

## The median

Another useful measure of the centre of a distribution of a numerical variable is the middle value, or **median**. To find the value of the median, all the observations are listed in order and the middle one is the median.

For example, the median of the following data set is 6, as there are five observations on either side of this value when the data are listed in order.

median = 6
↓
2 3 4 5 5 **6** 7 7 8 8 11

When there is an even number of data values, the median is defined as the midpoint of the two middle values. For example, the median of the following data set is 6.5, as there are six observations on either side of this value when the data are listed in order.

median = 6.5
↓
2 3 4 5 5 **6 7** 7 8 8 11 11

Returning to the premiership data. As the data are already given in order, it only remains to determine the middle observation.

Since there are 18 entries in the table there is no actual middle observation, so the median is chosen as the value halfway between the two middle observations, in this case the ninth and tenth (5 and 4).

$$\text{median} = \frac{1}{2}(5 + 4) = 4.5$$

The interpretation here is that, of the teams in the AFL, half (or 50%) have won the premiership 5 or more times and half (or 50%) have won the premiership 4 or less times.

The following rule is useful for locating the median in a larger data set stem plot.

### Determining the median

To compute the median of a distribution:

- arrange all the observations in ascending order according to size
- if $n$, the number of observations, is odd, then the median is the $\frac{n+1}{2}$th observation from the end of the list
- if $n$, the number of observations, is even, then the median is found by averaging the two middle observations in the list. That is, to find the median the $\frac{n}{2}$ and the $\left(\frac{n}{2}+1\right)$th observations are added together and divided by 2.

### Example 13 Determining the median

Find the median age of 23 people whose ages are displayed in the ordered stem plot below.

Age (years) 1 | 2 represents 12 years

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

**Solution**

As the data are already given in order, it only remains to determine the middle observation.

**1** Write down the number of observations.

$n = 23$

**2** The median is located at the $\frac{n+1}{2}$th position.

median is at the $\frac{23+1}{2} = 12$th position

Thus the median age is 41 years.

Note: We can check to see whether we are correct by counting the number of data values either side of the median. They should be equal.

## Comparing the mean and median

In Example 12 we found that the mean number of premierships won by the 18 AFL clubs was $\bar{x} = 6.5$. By contrast, in Example 13, we found that the median number of premierships won was 4.5.

These two values are quite different and the interesting question is: Why are they different, and which is the better measure of centre in this situation?

To help us answer this question, consider a stem plot of these data values.

Premierships won

| | |
|---|---|
| 0 | 0 0 0 1 1 1 2 3 4 |
| 0 | 5 9 |
| 1 | 0 1 1 2 |
| 1 | 5 6 6 |

From the stem-and-leaf plot it can be seen that the distribution is positively skewed. This example illustrates a property of the mean. When the distribution is skewed or if there are one or two very extreme values, then the value of the mean may be far from the centre. The median is not so affected by unusual observations and always gives the middle value.

## ▶ Measures of spread

A measure of spread is calculated in order to judge the *variability* of a data set. That is, are most of the values clustered together, or are they rather spread out?

### The range

The simplest measure of spread can be determined by considering the difference between the smallest and the largest observations. This is called the **range**.

#### The range

The range ($R$) is the simplest measure of spread of a distribution.

The range is the difference between the largest and smallest values in the data set.

$R$ = largest data value − smallest data value

#### Example 14 Finding the range

Consider the marks, for two different tasks, awarded to a group of students:

Task A

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 6 | 9 | 10 | 11 | 12 | 13 | 22 | 23 | 24 | 26 | 26 | 27 | 33 | 34 |
| 35 | 38 | 38 | 39 | 42 | 46 | 47 | 47 | 52 | 52 | 56 | 56 | 59 | 91 | 94 |

Task B

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 16 | 19 | 21 | 23 | 28 | 31 | 31 | 33 | 38 | 41 | 49 | 52 | 53 | 54 |
| 56 | 59 | 63 | 65 | 68 | 71 | 72 | 73 | 75 | 78 | 78 | 78 | 86 | 88 | 91 |

Find the range of each of these distributions.

#### Solution

For task A the minimum mark is 2 and the maximum mark is 94.

Range for Task A = 94 − 2 = 92

For Task B, the minimum mark is 11 and the maximum mark is 91.

Range for Task B = 91 − 11 = 80

ISBN 978-1-107-56755-9

The range for Task A is greater than the range for Task B. Is the range a useful summary statistic for comparing the spread of the two distributions? To help make this decision, consider the stem plots of the data sets:

Task A

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

Task B

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

From the stem-and-leaf plots of the data it appears that the spread of marks for the two tasks is not really described by the range. It is clear that the marks for Task A are more concentrated than the marks for Task B, except for the two unusual values for Task A.

Another measure of spread is needed, one which is not so influenced by these extreme values. The statistic we use for this task is the **interquartile range**.

## The interquartile range

The interquartile range (IQR) gives the spread of the middle 50% of data values.

### 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, and $Q_3$, the *third quartile*, which is the median of the upper half of the observations.

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

Definitions of the **quartiles** of a distribution sometimes differ slightly from the one given here. Using different definitions may result in slight differences in the values obtained, but these will be minimal and should not be considered a difficulty.

ISBN 978-1-107-56755-9

Cambridge University Press

### Example 15 Finding the interquartile range (IQR)

Find the interquartile ranges for Tasks A and B in Example 14 and compare.

**Solution**

1 There are 30 values in total. This means that there are fifteen values in the lower 'half', and fifteen in the upper 'half'. The median of the lower half ($Q_1$) is the 8th value.

Task A

Lower half:

2 6 9 10 11 12 13 22 23 24 26 26 27 33 34

$Q_1 = 22$

2 The median of the upper half ($Q_3$) is the 8th value.

Upper half:

35 38 38 39 42 46 47 47 52 52 56 56 59 91 94

$Q_3 = 47$

3 Determine the IQR.

$\text{IQR} = Q_3 - Q_1 = 47 - 22 = 25$

4 Repeat the process for Task B.

Task B

$Q_1 = 31$

$Q_3 = 73$

$\text{IQR} = Q_3 - Q_1 = 73 - 31 = 42$

5 Compare the IQR for Task A to the IQR for Task B.

The IQR shows the variability of Task A marks is smaller than the variability of Task B marks.

The interquartile range describes the range of the middle 50% of the observations. It measures the spread of the data distribution around the median ($M$). Since the upper 25% and the lower 25% of the observations are discarded, the interquartile range is generally not affected by outliers in the data set, which makes it a reliable measure of spread.

## The standard deviation

The **standard deviation** (*s*), measures the spread of a data distribution about the mean ($\bar{x}$).

### The standard deviation

The standard deviation is defined to be:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

where $n$ is the number of data values (sample size) and $\bar{x}$ is the mean.

Although it is 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 will always be zero. For technical reasons we average by dividing by $n - 1$, not $n$. In practice this is not a problem, as dividing by $n - 1$ compared to $n$ generally makes very little difference to the final value.

Normally, you will use your calculator to determine the value of a standard deviation. However, to understand what is involved when your calculator is doing the calculation, you should know how to calculate the standard deviation from the formula.

### Example 16 Calculating the standard deviation

Use the formula:

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

to calculate the standard deviation of the data set: 2, 3, 4.

**Solution**

1 To calculate $s$, it is convenient to set up a table with columns for:
$x$ the data values
$(x-\bar{x})$ the deviations from the mean
$(x-\bar{x})^2$ the squared deviations.

| | $x$ | $(x-\bar{x})$ | $(x-\bar{x})^2$ |
|---|---|---|---|
| | 2 | −1 | 1 |
| | 3 | 0 | 0 |
| | 4 | 1 | 1 |
| Sum | 9 | 0 | 2 |

2 First find the mean and then complete the table as shown.

$$\bar{x} = \frac{\sum x}{n} = \frac{2+3+4}{3} = \frac{9}{3} = 3$$

3 Substitute the required values into the formula and evaluate.

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{2}{3-1}} = 1$$

## ▶ Using a CAS calculator to calculate summary statistics

As you can see, calculating the various summary statistics you have encountered in this section is sometimes rather complicated and generally time consuming. Fortunately, it is no longer necessary to carry out these computations by hand, except in the simplest cases.

### How to calculate measures of centre and spread using the TI-Nspire CAS

The table shows the monthly rainfall figures for a year in Melbourne.

| *Month* | J | F | M | A | M | J | J | A | S | O | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Rainfall (mm)* | 48 | 57 | 52 | 57 | 58 | 49 | 49 | 50 | 59 | 67 | 60 | 59 |

Determine the mean and standard deviation, median and interquartile range, and range.

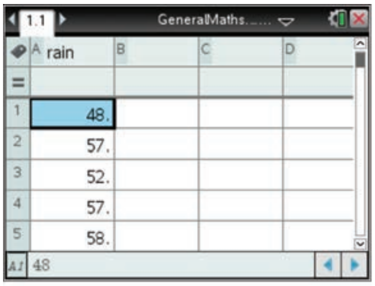
ISBN 978-1-107-56755-9

Cambridge University Press

### Steps

**1** Start a new document: Press [on] and select **New Document** (or press [ctrl] + [N]).

**2** Select **Add Lists & Spreadsheet**.
Enter the data into a list named *rain* as shown.
Statistical calculations can be done in the **Lists & Spreadsheet** application or the **Calculator** application.



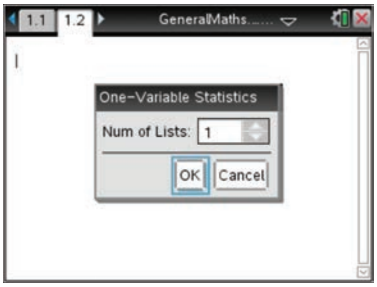
| | A rain | B | C | D |
|---|---|---|---|---|
| = | | | | |
| 1 | 48. | | | |
| 2 | 57. | | | |
| 3 | 52. | | | |
| 4 | 57. | | | |
| 5 | 58. | | | |

A1 48

**3** Press [ctrl] + [I] and select **Add Calculator** (or press [on] and arrow to [× ÷ + −] and press [enter]).

**a** Press [menu]**>Statistics>Stat Calculations> One-Variable Statistics**, [enter].

**b** Press the [tab] key to highlight OK and [enter].

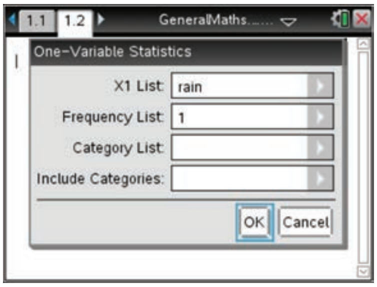


One-Variable Statistics

Num of Lists: 1

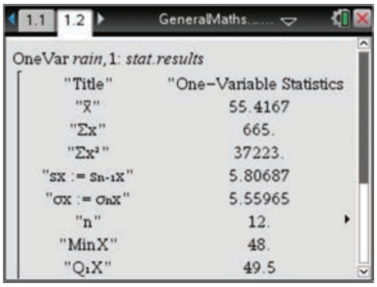
**c** Use the ▶ arrow and [enter] to paste in the list name **rain**. Press [esc] to exit the popup screen and generate statistical results screen below.

Notes: 1 The sample standard deviation is **sx**.
2 Use the ▲ arrows to scroll through the results screen to see the full range of statistics calculated.

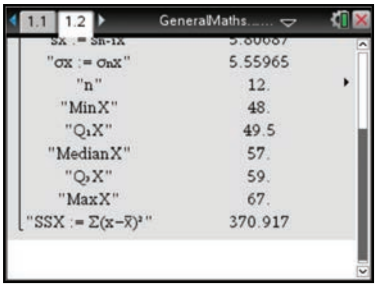


One-Variable Statistics

X1 List: rain
Frequency List: 1
Category List:
Include Categories:



```
OneVar rain,1: stat.results
   "Title"          "One-Variable Statistics
   "x̄"              55.4167
   "Σx"             665.
   "Σx²"            37223.
 "sx := sn-1x"      5.80687
 "σx := σnx"        5.55965
   "n"              12.
   "MinX"           48.
   "Q₁X"            49.5
```



```
  "sx := sn-1x"     5.80687
  "σx := σnx"       5.55965
      "n"           12.
     "MinX"         48.
     "Q₁X"          49.5
   "MedianX"        57.
     "Q₃X"          59.
     "MaxX"         67.
 "SSX := Σ(x-x̄)²"   370.917
```

**4** Write the answers correct to one decimal place.

$\bar{x} = 55.4, s = 5.8$

$M = 57$

$IQR = Q_3 - Q_1 = 59 - 49.5 = 9.5$

$R = max - min = 67 - 48 = 19$

ISBN 978-1-107-56755-9

Cambridge University Press

## How to calculate measures of centre and spread using the ClassPad

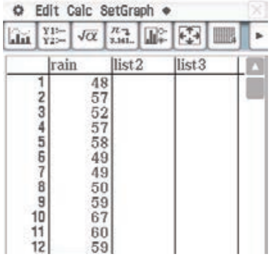
The table shows the monthly rainfall figures for a year in Melbourne.

| *Month* | J | F | M | A | M | J | J | A | S | O | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Rainfall (mm)* | 48 | 57 | 52 | 57 | 58 | 49 | 49 | 50 | 59 | 67 | 60 | 59 |

Determine the mean and standard deviation, median and interquartile range, and range.

### Steps

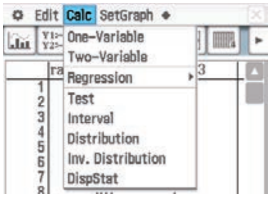
**1** Open the **Statistics** application and enter the data into the column labelled **rain**.



```
 1  48
 2  57
 3  52
 4  57
 5  58
 6  49
 7  49
 8  50
 9  59
10  67
11  60
12  59
```

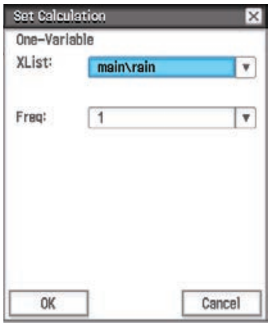
**2** To calculate the mean, median, standard deviation, and quartiles:

- Select **Calc** from the menu bar.
- Tap **One-Variable** and open the **Set Calculation** dialog box.



**3** Complete the dialog box. For:

- **XList:** select **main \ rain** (▼)
- **Freq:** leave as **1**.

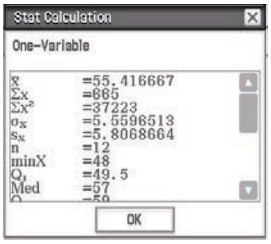


One-Variable
XList: main\rain
Freq: 1

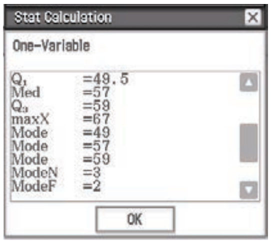
**4** Tap **OK** to confirm your selections.

Notes: 1 The sample standard deviation is given by $s_x$.

2 Use the ▲▼ side-bar arrows to scroll through the results screen to obtain values for additional statistics if required.



```
One-Variable
x̄       =55.416667
Σx      =665
Σx²     =37223
σx      =5.5596513
sx      =5.8068664
n       =12
minX    =48
Q1      =49.5
Med     =57
Q3      =59
```



```
One-Variable
Q1      =49.5
Med     =57
Q3      =59
maxX    =67
Mode    =49
Mode    =57
Mode    =59
ModeN   =3
ModeF   =2
```

**5** Write the answers correct to one decimal place.

$\bar{x} = 55.4, s = 5.8$

$M = 57$

$IQR = Q_3 - Q_1 = 59 - 49.5 = 9.5$

$R = \max - \min = 67 - 48 = 19$

# Exercise 2G

## Calculating the mean, median and IQR without a calculator

Example 12 **1** Find, without using a calculator, the mean and median for each of these data sets.

**a** 2 5 7 2 9

**b** 4 11 3 5 6 1

**c** 15 25 10 20 5

**d** 101 105 98 96 97 109

**e** 1.2 1.9 2.3 3.4 7.8 0.2

Example 14 **2** Find, without using a calculator, the median and IQR and range of each of these ordered data sets.

Example 15

**a** 2 2 5 7 9 11 12 16 23

**b** 1 3 3 5 6 7 9 11 12 12

**c** 21 23 24 25 27 27 29 31 32 33

**d** 101 101 105 106 107 107 108 109

**e** 0.2 0.9 1.0 1.1 1.2 1.2 1.3 1.9 2.1 2.2 2.9

Example 13 **3** Without a calculator, determine the median and the IQR for the data displayed in the following stem plots.

**a** Monthly rainfall (mm)

```
4 | 8 9 9
5 | 0 2 7 7 8 9 9
6 | 0 7
```

**b** Battery time (hours)

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

## Using a calculator to determine summary statistics

**4** The following table gives the area, in hectares, of each of the suburbs of a city:

3.6 2.1 4.2 2.3 3.4 40.3 11.3 19.4 28.4 27.6 7.4 3.2 9.0

**a** Find the mean and the median areas.

**b** Which is a better measure of centre for this data set? Explain your answer.

**5** The prices, in dollars, of apartments sold in a particular suburb during one month were:

$387 500 $329 500 $293 400 $600 000 $318 000 $368 000 $750 000
$333 500 $335 500 $340 000 $386 000 $340 000 $404 000 $322 000

**a** Find the mean and the median of the prices.

**b** Which is a better measure of centre of this data set? Explain your answer.

ISBN 978-1-107-56755-9

Cambridge University Press

Example 16

**6** A manufacturer advertised that a can of soft drink contains 375 mL of liquid. A sample of 16 cans yielded the following contents:

357 375 366 360 371 363 351 369
358 382 367 372 360 375 356 371

Find the mean and standard deviation, median and IQR, and range for the volume of drink in the cans. Give answers correct to one decimal place.

**7** The serum cholesterol levels for a sample of 20 people are:

231 159 203 304 248 238 209 193 225 244
190 192 209 161 206 224 276 196 189 199

Find the mean and standard deviation, median and IQR, and range of the serum cholesterol levels. Give answers correct to one decimal place.

**8** Twenty babies were born at a local hospital on one weekend. Their birth weights are given in the stem plot.

Birth weight (kg) 3 | 6 represents 3.6 kg

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

Find the mean and standard deviation, median and IQR, and range of the birth weights.

**9** The results of a student's chemistry experiment were as follows:

7.3 8.3 5.9 7.4 6.2 7.4 5.8 6.0

**a** **i** Find the mean and the median of the results.

**ii** Find the IQR and the standard deviation of the results.

**b** Unfortunately, when the student was transcribing his results into his chemistry book, he made a small error and wrote:

7.3 8.3 5.9 7.4 6.2 7.4 5.8 60

**i** Find the mean and the median of these results.

**ii** Find the interquartile range and the standard deviation of these results.

**c** Describe the effect the error had on the summary statistics in parts **a** and **b**.

ISBN 978-1-107-56755-9

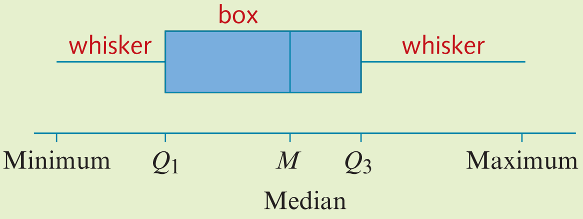
Cambridge University Press

## 2H Boxplots

Knowing the median and quartiles of a distribution means that quite a lot is known about the central region of the data set. If something is known about the tails of the distribution as well, then a good picture of the whole data set can be obtained. This can be achieved by knowing the **maximum** and **minimum** values of the data.

When we list the *median*, the *quartiles* and the *maximum* and *minimum* values of a data sets, we have what is known as a **five-number summary**. Its pictorial (graphical) representation is called a **boxplot** or a box-and-whisker plot.

### Boxplots



- A boxplot is a graphical representation of a five-number summary.
- A box is used to represent the middle 50% of scores.
- The median is shown by a vertical line drawn within the box.
- Lines (whiskers) extend out from the lower and upper ends of the box to the smallest and largest data values of the data set respectively.

### Example 17 Constructing a boxplot from a five-number summary

The following are the monthly rainfall figures for a year in Melbourne.

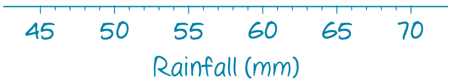
| *Month* | J | F | M | A | M | J | J | A | S | O | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Rainfall (mm)* | 48 | 57 | 52 | 57 | 58 | 49 | 49 | 50 | 59 | 67 | 60 | 59 |

Construct a boxplot to display this data, given the five-number summary:

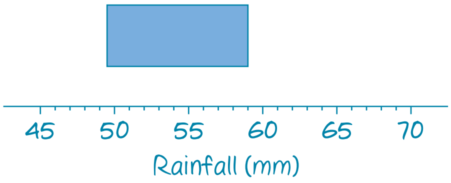
Min = 48, $Q_1$ = 49.5, $M$ = 57, $Q_3$ = 59, Max = 67

### Solution

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



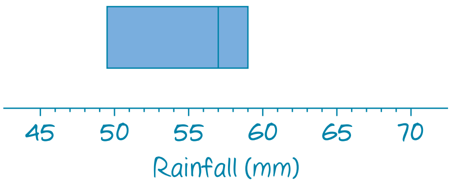
**2** Draw in a box starting at $Q_1$ = 49.5 and ending at $Q_3$ = 59.



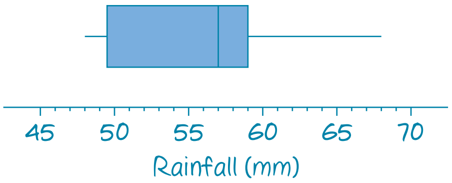
ISBN 978-1-107-56755-9

Cambridge University Press

**3** Mark in the median value with a vertical line segment at $M = 57$.



**4** Draw in the whiskers, lines joining the midpoint of the ends of the box, to the minimum and maximum values, 48 and 67, respectively.



## Boxplots with outliers

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

### Outlier

An *outlier* is a data value that appears to be rather different from other observations.

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?

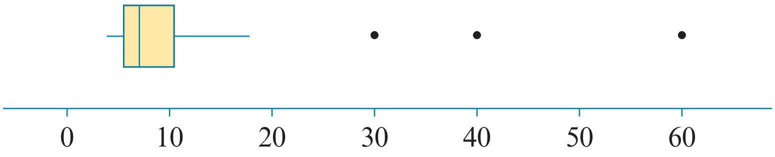
- One explanation is that the data distribution is extremely skewed with lots of data values in its tail.
- Another explanation is that the long whisker hides one or more outliers.

By modifying the boxplots, we can decide which explanation is most likely.

### Designating outliers

Any data point in a distribution that lies more than 1.5 interquartile ranges above the third quartile or more than 1.5 interquartile ranges below the first quartile could be an outlier.

These data values are plotted individually in the boxplot, and the whisker now ends at the largest or smallest data value that is not outside these limits. An example of a boxplot displaying outliers is shown below.



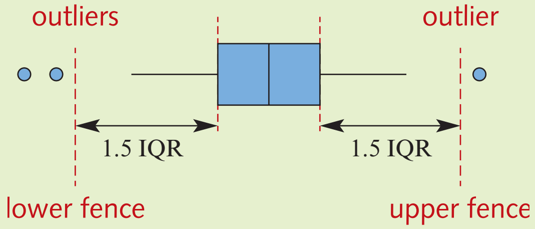
### Upper and lower fences

When constructing a boxplot to display outliers, we must first determine the location of what we call the *upper and lower fences*. These are imaginary lines drawn one and a half the interquartile range (or box widths) above and below the ends of the box (see over page). Data values outside these fences are classified as possible outliers and plotted separately.

### Using a boxplot to display possible outliers

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

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



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

## Example 18 Constructing a boxplot showing outliers

The number of hours that each of 33 students spent on a school project is shown below.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 3 | 4 | 9 | 9 | 13 | 19 | 24 | 27 | 35 | 36 |
| 37 | 40 | 48 | 56 | 59 | 71 | 76 | 86 | 90 | 92 | 97 |
| 102 | 102 | 108 | 111 | 146 | 147 | 147 | 166 | 181 | 226 | 264 |

Construct a boxplot for this data set that can be used to identify possible outliers.

### Solution

**1** From the ordered list, state the minimum and maximum values. Find the median, the $\frac{1}{2}(33 + 1)$th = 17th value.

min. = 2, max. = 264,
median = 71

**2** Determine $Q_1$ and $Q_3$. There are 33 values, so $Q_1$ is halfway between the 8th and 9th values and $Q_3$ is halfway between the 25th and the 26th values.

first quartile, $Q_1 = \dfrac{24 + 27}{2} = 25.5$

third quartile, $Q_3 = \dfrac{108 + 111}{2} = 109.5$

**3** Determine the IQR.

$\text{IQR} = Q_3 - Q_1 = 109.5 - 25.5 = 84$

**4** Determine the upper and lower fences.

$$\begin{aligned}\text{Lower fence} &= Q_1 - 1.5 \times \text{IQR}\\ &= 25.5 - 1.5 \times 84\\ &= -100.5\end{aligned}$$

$$\begin{aligned}\text{Upper fence} &= Q_3 + 1.5 \times \text{IQR}\\ &= 109.5 + 1.5 \times 84\\ &= 235.5\end{aligned}$$

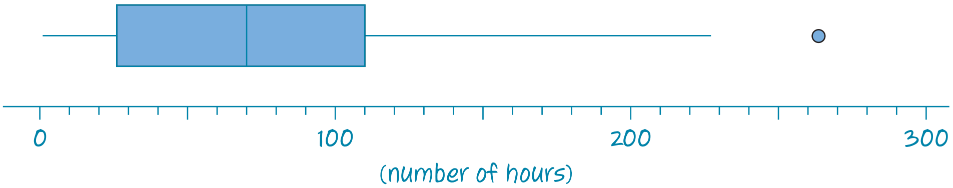
ISBN 978-1-107-56755-9

Cambridge University Press

5 Locate any values outside the fences, and the values that lie just inside the limits (the whiskers will extend to these values).

*There is one outlier 264.*
*The largest value that is not an outlier is 226.*

6 The boxplot can now be constructed as shown below. The circle denotes the outlier.



*There is one possible outlier, the student who spent 264 hours on the project.*

It is clearly very time-consuming to construct boxplots displaying outliers by hand. Fortunately, your CAS calculator will do it for you automatically as we will see below.

## How to construct a boxplot using the TI-Nspire CAS

The number of hours that each of 33 students spent on a school project is shown below.

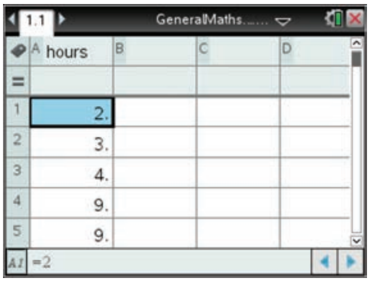
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 3 | 4 | 9 | 9 | 13 | 19 | 24 | 27 | 35 | 36 |
| 37 | 40 | 48 | 56 | 59 | 71 | 76 | 86 | 90 | 92 | 97 |
| 102 | 102 | 108 | 111 | 146 | 147 | 147 | 166 | 181 | 226 | 264 |

Construct a boxplot for this data set that can be used to identify possible outliers.

### Steps

1 Press on and select **New Document** (or use ctrl + N).

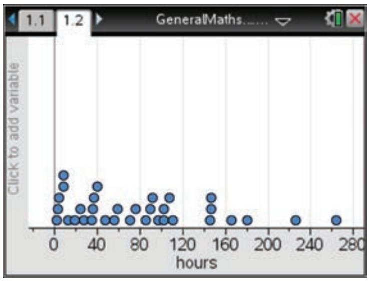
2 Select **Add Lists & Spreadsheet**.
Enter the data into a list called **hours** as shown.



| | A hours | B | C | D |
|---|---|---|---|---|
| = | | | | |
| 1 | 2. | | | |
| 2 | 3. | | | |
| 3 | 4. | | | |
| 4 | 9. | | | |
| 5 | 9. | | | |

A1 =2

3 Statistical graphing is done through the **Data & Statistics** application. Press ctrl + I and select **Add Data & Statistics** (or press on, arrow to , and press enter).
Note: A random display of dots will appear – this is to indicate list data is available for plotting. It is not a statistical plot.

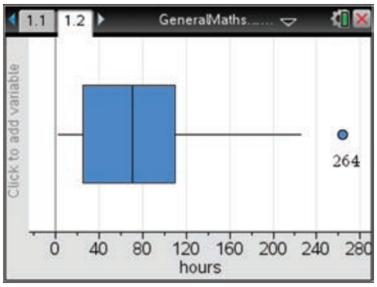


a Press tab to show the list of variables. Select the variable **hours**. Press enter to paste the variable **hours** to that axis. A dot plot is displayed as the default plot.

b To change the plot to a boxplot press menu>**Plot Type>BoxPlot**, then enter or click' (press ). Outliers are indicated by a dot(s).

ISBN 978-1-107-56755-9

Cambridge University Press



**4** Data Analysis

Move the cursor over the plot to display the key values (or use menu>**Analyze>Graph Trace**).

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

- minimum value is 2: **minX** = **2**
- first quartile is 25.5: $\mathbf{Q_1}$ = **25.5**
- median is 71: **Median** = **71**
- third quartile is 109.5: $\mathbf{Q_3}$ = **109.5**
- maximum value is 264: **maxX** = **264**. It is also an outlier.

## How to construct a boxplot using the ClassPad

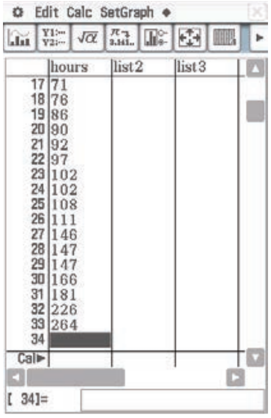
The number of hours that each of 33 students spent on a school project is shown below.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 3 | 4 | 9 | 9 | 13 | 19 | 24 | 27 | 35 | 36 |
| 37 | 40 | 48 | 56 | 59 | 71 | 76 | 86 | 90 | 92 | 97 |
| 102 | 102 | 108 | 111 | 146 | 147 | 147 | 166 | 181 | 226 | 264 |

Construct a box plot for this data set that can be used to identify possible outliers.

### Steps

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



| | hours | list2 | list3 |
|---|---|---|---|
| 17 | 71 | | |
| 18 | 76 | | |
| 19 | 86 | | |
| 20 | 90 | | |
| 21 | 92 | | |
| 22 | 97 | | |
| 23 | 102 | | |
| 24 | 102 | | |
| 25 | 108 | | |
| 26 | 111 | | |
| 27 | 146 | | |
| 28 | 147 | | |
| 29 | 147 | | |
| 30 | 166 | | |
| 31 | 181 | | |
| 32 | 226 | | |
| 33 | 264 | | |
| 34 | | | |

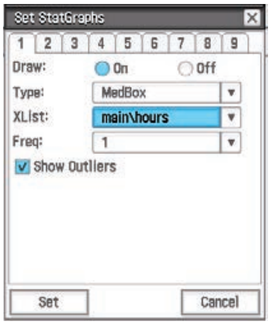
[ 34]=

**2** Open the **Set StatGraphs** dialog box by tapping the toolbar icon in the toolbar. Complete the dialog box as shown, right. For:

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

Tap the **Show Outliers** box.

Tap Set to exit.



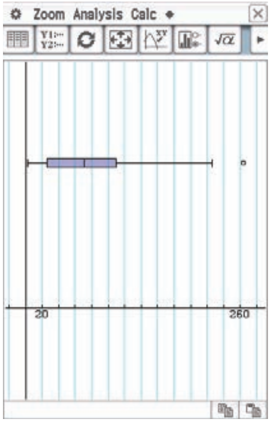
Draw: ● On ○ Off
Type: MedBox
XList: main\hours
Freq: 1
☑ Show Outliers

ISBN 978-1-107-56755-9

Cambridge University Press

**3** Tap 📊 to plot the boxplot.

**4** Tap Resize to obtain a full-screen display.

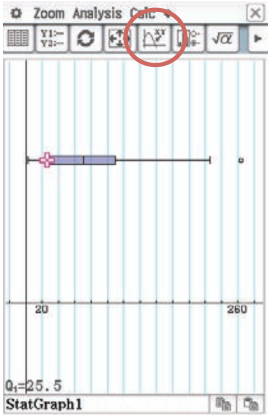


**5** Key values can be read from the boxplot by tapping 📈.

Use the arrows (◀ and ▶) to move from point to point on the boxplot.

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

- minimum value is 2 (**minX = 2**)
- first quartile is 25.5 (**$Q_1$ = 25.5**)
- median is 71 (**Median = 71**)
- third quartile is 109.5 (**$Q_3$ = 109.5**)
- maximum value is 264 (**maxX = 264**). It is also an outlier.



$Q_1$=25.5
StatGraph1

## Exercise 2H

### Constructing a boxplot from a five-number summary

Example 17

**1** The heights (in centimetres) of a class of girls are:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 160 | 165 | 123 | 143 | 154 | 180 | 133 | 123 | 157 | 157 |
| 135 | 140 | 140 | 150 | 154 | 159 | 149 | 167 | 176 | 163 |
| 154 | 167 | 168 | 132 | 145 | 143 | 157 | 156 | | |

The five-number summary for this data is:

Min = 123, $Q_1$ = 141.5, $M$ = 154, $Q_3$ = 161.5, Max = 180

Use this five-number summary to construct a boxplot (there are no outliers).

ISBN 978-1-107-56755-9

Cambridge University Press

**2** The data shows how many weeks each of the singles in the Top 41 has been in the charts, in a particular week.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 24 | 11 | 5 | 7 | 4 | 15 | 13 | 4 | 12 | 14 | 3 | 12 | 4 | 4 |
| 3 | 10 | 17 | 8 | 6 | 2 | 18 | 15 | 5 | 6 | 9 | 14 | 4 | 5 |
| 14 | 12 | 16 | 11 | 6 | 7 | 12 | 4 | 16 | 2 | 8 | 10 | 1 | |

The five-number summary for this data is:

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

Use this five-number summary to construct a boxplot (there are no outliers).

Example 18

**3** The amount of pocket money paid per week to a sample of year 8 students is:

| | | | | | | |
|---|---|---|---|---|---|---|
| \$5.00 | \$10.00 | \$12.00 | \$8.00 | \$7.50 | \$12.00 | \$15.00 |
| \$10.00 | \$10.00 | \$0.00 | \$5.00 | \$10.00 | \$20.00 | \$15.00 |
| \$26.00 | \$13.50 | \$15.00 | \$5.00 | \$15.00 | \$25.00 | \$16.00 |

The five-number summary for this data is:

Min = 0, $Q_1 = 7.75$, $M = 12$, $Q_3 = 15$, Max = 26

Use this five number summary to construct a boxplot (there is one outlier).

## Constructing boxplots from raw data

**4** The length of time, in years, that employees have been employed by a company is:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 20 | 8 | 6 | 9 | 13 | 15 | 4 | 2 |
| 15 | 14 | 13 | 4 | 16 | 18 | 26 | 6 | 8 | 2 |
| 6 | 7 | 20 | 2 | 1 | 1 | 5 | 8 | | |

Use a CAS calculator to construct the boxplot.

**5** The times (in seconds) that 35 children took to tie up a shoelace are:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 6 | 18 | 39 | 7 | 10 | 5 | 8 | 6 | 14 | 11 | 10 |
| 8 | 35 | 6 | 6 | 14 | 15 | 6 | 7 | 6 | 5 | 8 | 11 |
| 8 | 15 | 8 | 8 | 7 | 8 | 8 | 6 | 29 | 5 | 7 | |

Use a CAS calculator to construct the boxplot.

**6** A researcher is interested in the number of books people borrow from a library. She selected a sample of 38 people and recorded the number of books each person had borrowed in the previous year. Here are her results:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 28 | 0 | 2 | 38 | 18 | 0 | 0 | 4 | 0 | 0 | 5 | 13 |
| 2 | 13 | 1 | 1 | 14 | 1 | 8 | 27 | 0 | 52 | 4 | 11 | 0 |
| 0 | 12 | 28 | 15 | 10 | 1 | 0 | 2 | 0 | 1 | 11 | 0 | |

**a** Use a CAS calculator to construct a boxplot of the data.

**b** Use the boxplot to identify any possible outliers and write down their values.

**7** The following table gives the prices for units sold in a particular suburb in one month (in thousands of dollars):

| | | | | |
|---|---|---|---|---|
| 356 | 366 | 375 | 389 | 432 |
| 445 | 450 | 450 | 495 | 510 |
| 549 | 552 | 579 | 585 | 590 |
| 595 | 625 | 725 | 760 | 880 |
| 940 | 950 | 1017 | 1180 | 1625 |

**a** Use a CAS calculator to construct a boxplot of the data.

**b** Use the boxplot to identify any possible outliers and write down their values.

**8** The time taken, in seconds, for a group of children to complete a puzzle is:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 6 | 18 | 39 | 7 | 10 | 5 | 8 | 6 | 14 | 11 | 5 |
| 10 | 8 | 60 | 6 | 6 | 14 | 15 | 6 | 7 | 6 | 5 | 7 |
| 8 | 11 | 8 | 15 | 8 | 8 | 7 | 8 | 8 | 6 | 29 | |

**a** Use a CAS calculator to construct a boxplot of the data.

**b** Use the boxplot to identify any possible outliers and write down their values.

**9** The percentage of people using the internet in 23 countries is given in the table:

| Country | Internet users (%) |
|---|---|
| Afghanistan | 5.45 |
| Argentina | 55.80 |
| Australia | 79.00 |
| Brazil | 48.56 |
| Bulgaria | 51.90 |
| China | 42.30 |
| Colombia | 48.98 |
| Greece | 55.07 |
| Hong Kong SAR, China | 72.90 |
| Iceland | 96.21 |
| India | 12.58 |
| Italy | 55.83 |
| Malaysia | 65.80 |
| Morocco | 55.42 |
| New Zealand | 82.00 |
| Saudi Arabia | 54.00 |
| Singapore | 72.00 |
| Slovenia | 68.35 |
| South Africa | 41.00 |
| United Kingdom | 87.48 |
| United States | 79.30 |
| Venezuela | 49.05 |
| Vietnam | 39.49 |

**a** Use a CAS calculator to construct a boxplot of the data.

**b** Use the boxplot to identify any possible outliers and write down their values.

# 2I Comparing the distribution of a numerical variable across two or more groups

It makes sense to compare the distributions of data sets when they are concerned with the *same* numerical variable, say *height* measured for different groups of people, for example, a basketball team and a gymnastic team.

For example, it would be useful to compare the distributions for each of the following:

- the maximum daily temperatures in Melbourne in March and the maximum daily temperatures in Sydney in March
- the test scores for a group of students who had not had a revision class and the test scores for a group of students who had a revision class.

In each of these examples, we can actually identify *two variables*. One is a *numerical variable* and the other is a *categorical variable*.

For example:

- The variable maximum daily *temperature* is numerical while the variable *city*, which takes the values 'Melbourne' or 'Sydney', is categorical.
- The variable *test score* is numerical while the variable *attended a revision class*, which takes the values 'yes' or 'no', is categorical.

Thus, when we compare two data sets in this section, we will be actually investigating the relationship between two variables: a numerical variable and a categorical variable.

The outcome of these investigations will be a brief written report that compares the distribution of the numerical variable across two or more groups defined as categorical variables. The starting point for these investigations will be, as always, a graphical display of the data. To this end you will meet and learn to interpret two new graphical displays: the **back-to-back stem plot** and the **parallel boxplot**.

## ▶ Comparing distributions using back-to-back stem plots

A back-to-back stem plot differs from the stem plots you have met in the past in that it has a single stem with two sets of leaves, one for each of the two groups being compared.

ISBN 978-1-107-56755-9

Cambridge University Press

### Example 19 Comparing distributions using back-to-back stem plots

The following back-to-back stem plot displays the distributions of life expectancies for males (in years) in several countries in the years 1970 and 2010.

In this situation, *life expectancy* is the numerical variable. *Year*, which takes the values 1970 and 2010, is the categorical variable.

Male life expectancy (in years)

5|8 = 58 years

| 1970 | | 2010 |
|---:|:-:|:---|
| 8 | 3 | |
| | 4 | |
| 9 4 2 | 5 | 8 |
| 9 9 9 8 7 7 1 0 | 6 | 9 |
| 4 | 7 | 1 2 4 4 6 8 9 9 |
| | 8 | 0 0 0 |

| 1970 | 2010 |
|---|---|
| M = 67 years | M = 76 years |
| IQR = 12.5 years | IQR = 8 years |

Use the back-to-back stem plot and the summary statistics provided to compare these distributions in terms of centre and spread and draw an appropriate conclusion.

**Solution**

**1** Centre: Use the medians to compare centres.

The median life expectancy of males in 2010 (M = 76 years) was nine years higher than in 1970 (M = 67 years).

**2** Spread: Use the IQRs to compare spreads.

The spread of life expectancies of males in 2010 (IQR = 12.5 years) was different to the spread in 1970 (IQR = 8).

**3** Conclusion: Use the above observations to write a general conclusion.

In conclusion, the median life expectancy for these countries has increased over the last 40 years, and the variability in life expectancy between countries has decreased.

## ▶ Comparing distributions using parallel boxplots

Back-to-back stem plots can be used to compare the distribution of a numerical variable across two groups when the data sets are small. Parallel boxplots can also be used to compare distributions. Unlike back-to-back stem plots, boxplots can also be used when there are more than two groups.

By drawing boxplots on the same axis, both the centre and spread for the distributions are readily identified and can be compared visually.

ISBN 978-1-107-56755-9

Cambridge University Press

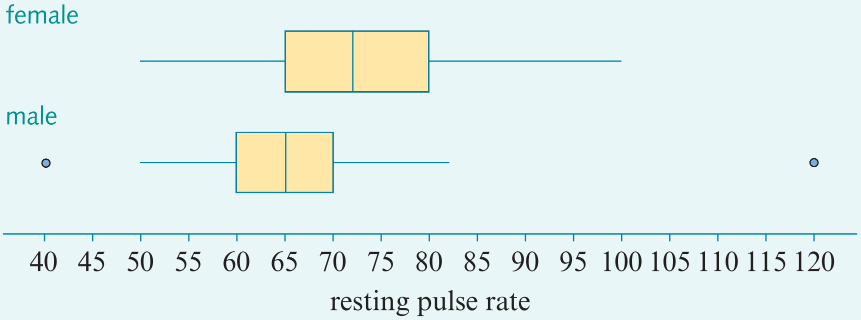
When comparing distributions of a numerical variable across two or more groups using parallel boxplots, the report should address the key features of:

- centre (the median)
- spread (the IQR)
- possible outliers.

### Example 20 Comparing distributions across two groups using parallel boxplots

The following parallel boxplots display the distribution of pulse rates (in beats/minute) for a group of female students and a group of male students.

Use the information in the boxplots to write a report comparing these distributions in terms of centre, spread and outliers in the context of the data.



**Solution**

**1** Centre: Compare the medians. Estimate values of these medians from the plot (the vertical lines in the boxes).

The median pulse rate for females ($M = 72$ beats/minute) is higher than that for males ($M = 65$ beats/minute).

**2** Spread: Compare the spread of the two distributions using IQRs (the widths of the boxes).

The spread of pulse rates for females ($IQR = 15$) is higher than for males ($IQR = 10$).

**3** Outliers: Locate on any outliers and describe.

There are no female outliers. The males with pulse rates of 40 and 120 were outliers.

**4** Conclusion: Use the above observations to write a general conclusion.

In conclusion, the median pulse rate for females was higher than for males and female pulse rates were generally more variable than male pulse rates.

ISBN 978-1-107-56755-9

## Exercise 2I

### Comparing groups using back-to-back stem plots

Example 19

**1** The stem plot displays the age distribution of ten females and ten males admitted to a regional hospital on the same day.

**a** Calculate the median and the IQR for admission ages of the females and males in this sample.

**b** Write a report comparing these distributions in terms of centre and spread in the context of the data.

| Females | | Males |
|---:|:---:|:---|
| 9 | 0 | |
| 5 0 | 1 | 3 6 |
| 7 | 2 | 1 4 5 6 7 |
| 7 1 | 3 | 4 |
| 3 0 | 4 | 0 7 |
| 0 | 5 | |
| | 6 | |
| 9 | 7 | |

4 | 0 = 40 years

**2** The stem plot opposite displays the mark distribution of students from two different mathematics classes (Class A and Class B) who sat the test. The test was marked out of 100.

**a** How many students in each class scored less than 50%?

**b** Determine the median and the IQR for the marks obtained by the students in each class.

**c** Write a report comparing these distributions in terms of centre and spread in the context of the data.

Marks %

| Class B | | Class A |
|---:|:---:|:---|
| 3 2 | 1 | 9 |
| | 2 | 2 |
| | 3 | 9 |
| | 4 | 5 7 8 |
| | 5 | 5 8 |
| 9 | 6 | 5 8 |
| 6 4 3 3 2 2 1 0 0 | 7 | 1 6 7 9 9 |
| 8 8 4 4 3 2 1 1 0 0 | 8 | 0 1 2 2 5 5 9 |
| 8 1 | 9 | 1 9 |

7 | 1 = 71

ISBN 978-1-107-56755-9

**3** The following table shows the number of nights spent away from home in the past year by a group of 21 Australian tourists and by a group of 21 Japanese tourists:

Australian

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | 14 | 15 | 3 | 6 | 17 | 2 |
| 7 | 4 | 8 | 23 | 5 | 7 | 21 |
| 9 | 11 | 11 | 33 | 4 | 5 | 3 |

Japanese

| | | | | | | |
|---|---|---|---|---|---|---|
| 14 | 3 | 14 | 7 | 22 | 5 | 15 |
| 26 | 28 | 12 | 22 | 29 | 23 | 17 |
| 32 | 5 | 9 | 23 | 6 | 44 | 19 |

**a** Construct a back-to-back stem-and-leaf plot of these data sets.

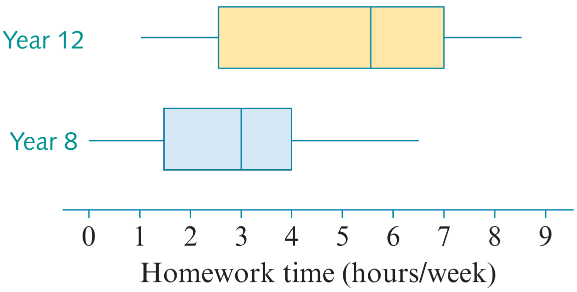
**b** Determine the median and IQR for the two distributions.

**c** Write a report comparing the distributions of the number of nights spent away by Australian and Japanese tourists in terms of centre and spread.

## Comparing groups using parallel boxplots

Example 20

**4** The boxplots below display the distributions of homework time (in hours/week) of a sample of year 8 and a sample of year 12 students.



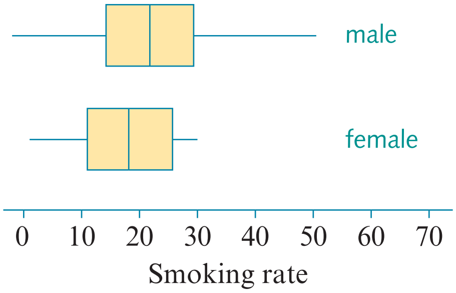
**a** Estimate the median and IQRs from the boxplots.

**b** Use these medians and IQRs to write a report comparing these distributions in terms of centre and spread in the context of the data.

**5** The boxplots below display the distribution of smoking rates (%) of males and females from several countries.

**a** Estimate the median and IQRs from the boxplots.

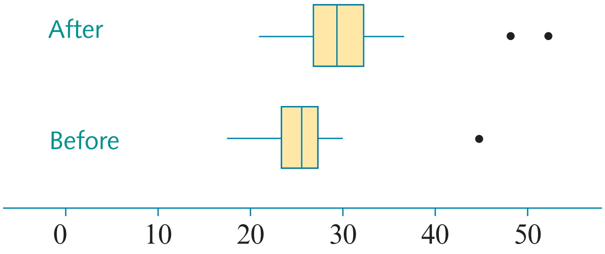
**b** Use the information in the boxplots to write a report comparing these distributions in terms of centre and spread in the context of the data.



ISBN 978-1-107-56755-9

Cambridge University Press

**6** The boxplots below display the distributions of the number of sit-ups a person can do in one minute, both before and after a fitness course.



**a** Estimate the median, IQRs and the values of any outliers from the boxplots.

**b** Use these medians and IQRs to write a report comparing these distributions in terms of centre and spread in the context of the data.

**7** To test the effect of alcohol on coordination twenty randomly selected participants were timed to complete a task with both 0% blood alcohol and 0.05% blood alcohol. The times taken (in seconds) are shown in the accompanying table.

| 0% blood alcohol | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 38 | 36 | 35 | 35 | 43 | 46 | 42 | 47 | 40 | 48 |
| 35 | 34 | 40 | 44 | 30 | 25 | 39 | 31 | 29 | 44 |

| 0.05% blood alcohol | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 39 | 32 | 35 | 39 | 36 | 34 | 41 | 64 | 44 | 38 |
| 43 | 42 | 46 | 46 | 50 | 32 | 32 | 41 | 40 | 50 |

**a** Draw boxplots for each of the sets of scores on the same scale.

**b** Use the information in the boxplots to write a report comparing the distributions of the times taken to complete a task with 0% blood alcohol and 0.05% blood alcohol in terms of centre (medians), spread (IQRs) and outliers.

# 2J Statistical investigation

## Exercise 2J

**1** To investigate the age of parents at the birth of their first child, a hospital recorded the ages of the mothers and fathers for the first 40 babies born in the hospital for each of the years 1970, 1990 and 2010.

The data is given below:

| 1970 Mother | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 23 | 22 | 33 | 19 | 19 | 26 | 20 | 15 | 26 | 17 |
| 18 | 31 | 24 | 20 | 29 | 28 | 25 | 45 | 28 | 22 |
| 1970 Father | | | | | | | | | |
| 29 | 15 | 39 | 29 | 22 | 35 | 32 | 26 | 37 | 29 |
| 25 | 31 | 20 | 34 | 28 | 22 | 33 | 25 | 34 | 46 |
| 1990 Mother | | | | | | | | | |
| 28 | 14 | 38 | 28 | 21 | 34 | 31 | 25 | 36 | 28 |
| 24 | 30 | 19 | 33 | 27 | 21 | 32 | 24 | 33 | 45 |
| 1990 Father | | | | | | | | | |
| 31 | 27 | 46 | 31 | 26 | 28 | 30 | 27 | 43 | 37 |
| 39 | 22 | 27 | 35 | 31 | 29 | 32 | 27 | 38 | 35 |
| 2010 Mother | | | | | | | | | |
| 30 | 26 | 45 | 32 | 25 | 27 | 29 | 26 | 42 | 36 |
| 38 | 21 | 26 | 34 | 37 | 28 | 28 | 37 | 37 | 34 |
| 2010 Father | | | | | | | | | |
| 37 | 31 | 39 | 36 | 21 | 34 | 34 | 23 | 17 | 37 |
| 23 | 33 | 31 | 32 | 24 | 39 | 45 | 30 | 35 | 34 |

Use appropriate displays and summary statistics to answer the following questions:

**a** How do the ages of the mothers compare to the ages of fathers in each time period?

**b** How have the ages of the mothers changed over the three time periods?

**c** How have the ages of the fathers changed over the three time periods?

**d** Has the relationship between mothers' ages and fathers' ages changed over time?

In each case write a brief report to summarise your findings.

ISBN 978-1-107-56755-9

## Key ideas and chapter summary

Types of data

Data can be classified as **numerical** or **categorical**.

Frequency table

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

**Frequency** can be recorded as the number of times a value occurs or a **percentage**, the percentage of times a value occurs.

Categorical data

**Categorical data** arises when classifying or naming some quality or attribute. When the categories are naming the groups, the data is called **nominal**. When there is an inherent order in the categories, the data is called **ordinal**.

Bar chart

A **bar chart** is used to display the frequency distribution of a categorical variable.

Mode, modal category/class

The **mode** (or modal category) is the value of a variable (or the category) that occurs most frequently. The **modal interval**, for **grouped data,** is the interval that occurs most frequently.

Numerical data

**Numerical data** arises from measuring or counting some quantity.

**Discrete** numerical data can only take particular values, usually whole numbers, and often arises from counting.

**Continuous** numerical data describes numerical data that can take any value, sometimes in an interval, and often arises from measuring.

Histogram

A **histogram** is used to display the frequency distribution of a numerical variable: suitable for medium to large-sized data sets.

Stem plot

A **stem plot** is a visual display of a numerical data set, an alternative display to the histogram: suitable for small to medium-sized data sets. Leading digits are shown as the stem and the final digit as the leaf.

Dot plot

A **dot plot** consists of a number line with each data point marked by a dot. Suitable for small to medium sized data sets.

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), **centre** (the midpoint of the distribution) and **spread**.

Summary statistics

**Summary statistics** are used to give numerical values to 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.

ISBN 978-1-107-56755-9

Cambridge University Press

**Range**

The **range** ***(R)*** is the difference between the smallest and the largest data values. It is the simplest measure of spread.

$$\text{range} = \text{largest value} - \text{smallest value}$$

**Standard deviation**

The **standard deviation** (*s*) is a summary statistic that measures the spread of the data values around the mean.

**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.
If the distribution is clearly skewed or there are outliers, the median is preferred to the mean as a measure of centre.

**Quartiles**

**Quartiles** are summary statistics that divide an ordered data set into four equal groups.

**Interquartile range**

The **interquartile range (IQR)** gives the spread of the middle 50% of data values in an ordered data set. If the distribution is highly skewed or there are outliers, the IQR is preferred to the standard deviation as a measure of spread.

**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.

**Boxplot**

A **boxplot** is a visual display of a five-number summary with adjustments made to display outliers separately when they are present.

## Skills check

Having completed this chapter you should be able to:

- differentiate between nominal, ordinal, discrete and continuous data
- interpret the information contained in a frequency table
- identify the mode from a frequency table and interpret it
- construct a bar chart or histogram from a frequency table
- construct a histogram from raw data using a graphics calculator
- construct a dot plot and stem-and-leaf plot from raw data
- recognise symmetric, positively skewed and negatively skewed distributions
- identify potential outliers in a distribution from its histogram or stem plot

- locate the median and quartiles of a data set and hence calculate the IQR
- produce a five-number summary from a set of data
- construct a boxplot from a five-number summary
- construct a boxplot from raw data using a graphics calculator
- use a boxplot to identify key features of a data set such as centre and spread
- use the information in a back-to-back stem plot or a boxplot to describe and compare distributions
- calculate the mean and standard deviation of a data set
- understand the difference between the mean and the median as measures of centre and be able to identify situations where it is more appropriate to use the median
- write a short paragraph comparing distributions in terms of centre, spread and outliers.

## Multiple-choice questions

**1** In a survey, a number of people were asked to indicate how much they exercised by selecting one of the options 'never', 'seldom', 'sometimes' or 'regularly'. The resulting variable was named *level of exercise*. The type of data generated is:

**A** variable **B** numerical **C** nominal **D** ordinal **E** metric

**2** For which of the following variables is a bar chart an appropriate display?

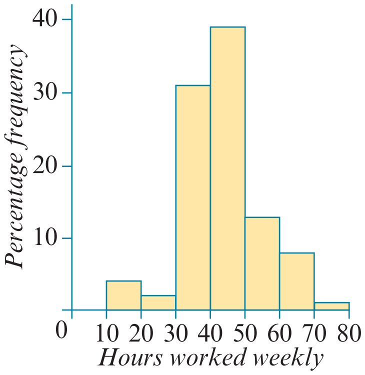
**A** Weight (kg) **B** Age (years)
**C** Distance between towns (km) **D** Hair colour
**E** Reaction time (seconds)

**3** For which of the following variables is a histogram an appropriate display?

**A** Hair colour **B** Sex (male, female)
**C** Distances between towns on a long road trip (km)
**D** Postcode **E** Weight (under weight, average, over weight)

*The following information relates to Questions 4 to 7*

The number of hours worked per week by employees in a large company is shown in the following percentage frequency histogram.



ISBN 978-1-107-56755-9

Cambridge University Press

**4** The percentage of employees who work from 20 to less than 30 hours per week is closest to:

**A** 1% **B** 2% **C** 6% **D** 10% **E** 33%

**5** The percentage of employees who worked *less* than 30 hours per week is closest to:

**A** 2% **B** 3% **C** 4% **D** 6% **E** 30%

**6** The modal interval for hours worked is:

**A** 10 to less than 20 **B** 20 to less than 30 **C** 30 to less than 40
**D** 40 to less than 50 **E** 50 to less than 60

**7** The median number of hours worked is in the interval:

**A** 10 to less than 20 **B** 20 to less than 30 **C** 30 to less than 40
**D** 40 to less than 50 **E** 50 to less than 60

*The following information relates to Questions 8 to 11*

A group of 18 employees of a company were asked to record the number of meetings they had attended in the last month.

1 1 2 3 4 5 5 6 7 9 10 12 14 14 16 22 23 44

**8** The range of meetings is:

**A** 22 **B** 23 **C** 24 **D** 43 **E** 44

**9** The median number of meetings is:

**A** 6 **B** 7 **C** 7.5 **D** 8 **E** 9

**10** The mean number of meetings is closest to:

**A** 7 **B** 8 **C** 9 **D** 10 **E** 11

**11** The interquartile range (IQR) of the number of meetings is:

**A** 0 **B** 4 **C** 9.5 **D** 10 **E** 14

**12** The heights of six basketball players (in cm) are:

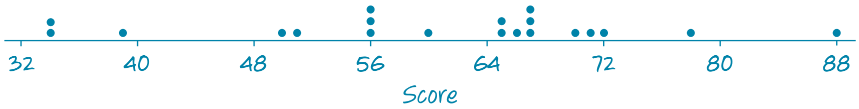
178.1 185.6 173.3 193.4 183.1 193.0

The mean and standard deviation are closest to:

**A** mean = 184.4; standard deviation = 8.0
**B** mean = 184.4; standard deviation = 7.3
**C** mean = 182.5; standard deviation = 7.3
**D** mean = 182.5; standard deviation = 8.0
**E** mean = 183.1; standard deviation = 7.3

*The following information relates to Questions 13 and 14*

The dot plot below gives the examination scores in mathematics for a group of 20 students.



**13** The number of students who scored 56 on the examination is:

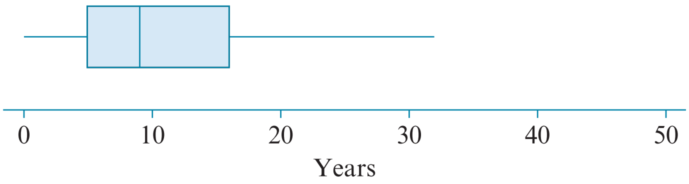
**A** 1 **B** 2 **C** 3 **D** 4 **E** 5

**14** The percentage of students who scored between 40 and 80 on the exam is closest to:

**A** 60% **B** 70% **C** 80% **D** 90% **E** 100%

*The following information relates to Questions 15 to 18*

The number of years for which a sample of people have lived at their current address is summarised in the boxplot.



**15** The range is closest to:

**A** 10 **B** 15 **C** 20

**D** 25 **E** 30

**16** The median number of *years lived at this address* is closest to:

**A** 5 **B** 9 **C** 12 **D** 15 **E** 47

**17** The interquartile range of the number of *years lived at this address* is closest to:

**A** 5 **B** 10 **C** 15 **D** 20 **E** 45

**18** The percentage who have lived at this address for more than 15 years is closest to:

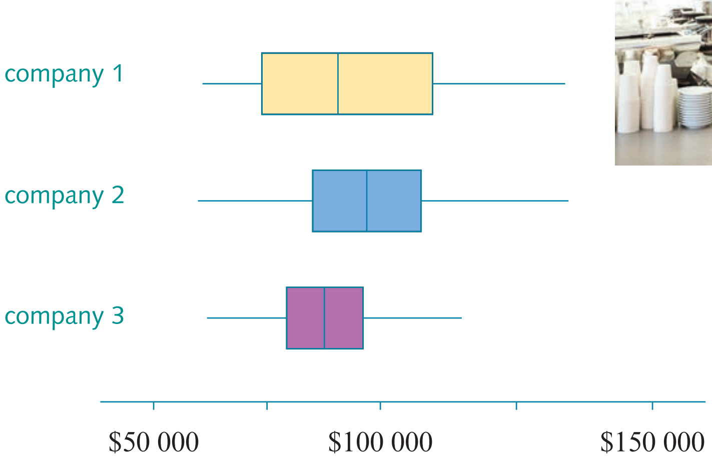
**A** 10% **B** 25% **C** 50% **D** 60% **E** 75%

ISBN 978-1-107-56755-9

Cambridge University Press

*The following information relates to Questions 19 to 21*

The amount paid per annum to the employees of each of three large companies is shown in the boxplots.



**19** The company with the lowest median wage is:

**A** company 1 **B** company 2 **C** company 3
**D** company 1 and company 2 **E** company 2 and company 3

**20** The company with the largest general spread (IQR) in wages is:

**A** company 1 **B** company 2 **C** company 3
**D** company 1 and company 2 **E** company 2 and company 3

**21** Which of the following statements is *not* true?

**A** All workers in company 3 earned less than $125 000 per year.

**B** More than half of workers in company 2 earned less than $100 000 per year.

**C** 75% of workers in company 2 earned less than the median wage in company 3.

**D** More than half of the workers in company 1 earned more than the median wage in company 3.

**E** More than 25% of the workers in company 1 earned more than the median wage at company 2.

ISBN 978-1-107-56755-9

Cambridge University Press

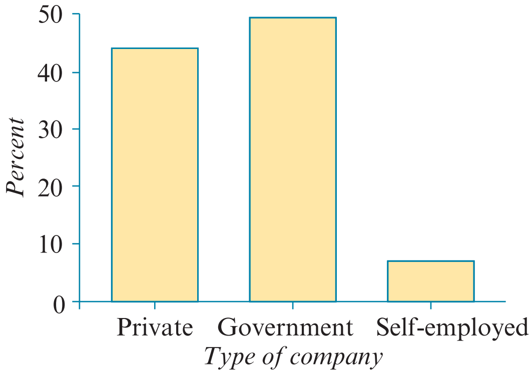
## Short-answer questions

**1** Classify the data that arises from the following situations as nominal, ordinal, discrete or continuous.

**a** The number of phone calls a hotel receptionist receives each day.

**b** Interest in politics on a scale from 1 to 5, where 1 = very interested, 2 = quite interested, 3 = somewhat interested, 4 = not very interested, and 5 = uninterested.

**2** The following bar chart shows the percentage of working people in a certain town who are employed in private companies, work for the government or are self-employed.



**a** Is the data categorical or numerical?

**b** Approximately what percentage of the people are self-employed?

**3** A researcher asked a group of people to record how many cigarettes they had smoked on a particular day. Here are her results:

0 0 9 10 23 25 0 0 34 32 0 0 30 0 4
5 0 17 14 3 6 0 33 23 0 32 13 21 22 6

Using class intervals of width 5, construct a histogram of this data.

**4** A teacher recorded the time taken (in minutes) by each of a class of students to complete a test:

56 57 47 68 52 51 43 22 59 51 39
54 52 69 72 65 45 44 55 56 49 50

**a** Make a dot plot of this data.

**b** Make a stem-and-leaf plot of these times.

**c** Use this stem plot to find the median and quartiles for the time taken.

**5** The monthly phone bills, in dollars, for a group of people are given below:

285 185 210 215 320 680 280
265 300 210 270 190 245 315

Find the mean and standard deviation, the median and the IQR, and the range of the monthly phone bills.

ISBN 978-1-107-56755-9

Cambridge University Press

**6** Geoff decided to record the time (in minutes) it takes him to complete his mail round each working day for four weeks. His data is recorded below:

170 189 201 183 168 182 161 166 167 173 182 167 188 211
164 176 161 187 180 201 147 188 186 176 174 193 185 183

Find the mean and standard deviation of his mail round times.

**7** A group of students was asked to record the number of SMS messages that they sent in one 24-hour period. The following five-number summary was obtained from the data set.

$$\text{Min} = 0, \qquad Q_1 = 3, \qquad M = 5, \qquad Q_3 = 12, \qquad \text{Max} = 24$$

Use the summary to construct a boxplot of this data.

**8** The following data gives the number of students absent from a large secondary college on each of 36 randomly chosen school days:

7 22 12 15 21 16 23 23 17 23 8 16
7 3 21 30 13 2 7 12 18 14 14 0
15 16 13 21 10 16 11 4 3 0 31 44

**a** Construct a boxplot of this data.
**b** What was the median number of students absent each day during this period?
**c** On what percentage of days were more than 20 students absent?

## Extended-response questions

**1** The divorce rates (in percentages) of 19 countries are:

27 18 14 25 28 6 32 44 53 0
26 8 14 5 15 32 6 19 9

**a** Is the data categorical or numerical?
**b** Construct an ordered stem plot of divorce rates by hand.
**c** Construct a dot plot of divorce rates by hand.
**d** What shape is the distribution of divorce rates?
**e** What percentage of the 19 countries have divorce rates greater than 30%?
**f** Calculate the mean and median of the distribution of divorce rates.
**g** Use your calculator to construct a histogram of the data with class intervals of width 10.

**i** What is the shape of the histogram?
**ii** How many of the 19 countries have divorce rates from 10% to less than 20%?

**2** Metro has decided to improve its service on the Lilydale line. Trains were timed on the run from Lilydale to Flinders Street, and their times recorded over a period of six weeks at the same time each day.
The journey times are shown below (in minutes):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 60 | 61 | 70 | 72 | 68 | 80 | 76 | 65 | 69 | 79 | 82 |
| 90 | 59 | 86 | 70 | 77 | 64 | 57 | 65 | 60 | 68 | 60 |
| 63 | 67 | 74 | 78 | 65 | 68 | 82 | 89 | 75 | 62 | 64 |
| 58 | 64 | 69 | 59 | 62 | 63 | 89 | 74 | 60 | | |

**a** Use your CAS calculator to construct a histogram of the times taken for the journey from Lilydale to Flinders Street.

**i** On how many days did the trip take 65–69 minutes?

**ii** What shape is the histogram?

**iii** What percentage of trains took less than 65 minutes to reach Flinders Street?

**b** Use your calculator to determine the following summary statistics for the *time* taken (correct to two decimal places):

$$\bar{x},\ s,\ \text{Min},\ Q_1,\ M, Q_3,\ \text{Max}$$

**c** Use the summary statistics to complete the following report.

**i** The mean time taken from Lilydale to Flinders Street was □ minutes.

**ii** 50% of the trains took more than □ minutes to travel from Lilydale to Flinders Street.

**iii** The range of travelling times was □ minutes, while the interquartile range was □ minutes.

**iv** 25% of trains took more than □ minutes to travel to Flinders Street.

**v** The standard deviation of travelling times was □ minutes.

**d** Summary statistics for the year before Metro took over the Lilydale line from Connex are:

$$\text{Min} = 55, \quad Q_1 = 65, \quad M = 70, \quad Q_3 = 89, \quad \text{Max} = 99$$

Construct boxplots for the last year Connex ran the line and for the data from Metro on the same plot.

**e** Use the information from the boxplots to write a report comparing the distribution of travelling times for the two transport corporations in terms of centre (medians) and spread (IQRs).