

Chapter 2

Investigating and comparing data distributions

Chapter questions

- ▶ 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 are dot and stem plots and when are they used?
- ▶ What are the mean, median, range, interquartile range and standard deviation?
- ▶ What is the five-number summary?
- ▶ What is an outlier?
- ▶ How do we construct and interpret boxplots?
- ▶ How do we use stem plots and boxplots to compare two or more groups?

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 help us to answer real-world questions when the data are collected from a **single variable**.

2A Classifying and displaying categorical data

Learning intentions

- ▶ To be able to classify data as categorical or numerical.
- ▶ To be able to further classify categorical data as nominal or ordinal.
- ▶ To be able to further classify numerical data as discrete or continuous.
- ▶ To be able to construct frequency and percentage frequency tables for categorical data.
- ▶ To be able to construct bar charts and percentage bar charts from frequency tables.
- ▶ To be able to interpret and describe frequency tables and bar charts.

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

- indicate their gender 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, and
- 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: *gender*, *coffee size*, *number of brothers* and *hand span*.

<i>Gender</i>	<i>Coffee size</i>	<i>Number of brothers</i>	<i>Hand span (in cm)</i>
M	Large	0	23.6
F	Small	2	19.6
F	Small	1	20.2
M	Large	1	24.0

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

The variables in the table, and hence the data collected from that variable, fall into two broad types: **categorical** or **numerical**.

Categorical data

The type of 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 gender 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 the values of the variable are simply names.

- 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 allows us to both name and order their responses.

Numerical data

The type of 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, ... Because the data can only take particular numerical values, it is called **discrete data**. Discrete data arises in situations where counting is involved.
- In response to the hand span question, students who wrote 24.1 cm could have an actual hand span of anywhere between 23.05 and 24.04 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.

A quite different distinction is sometimes made between numerical data which is **interval** and numerical data which is **ratio**.

- An **interval** scale requires only that the differences between successive steps on the scale are equal. The temperature scale is an example of an interval scale; the amount of heat needed to raise the temperature from 13°C to 14°C is the same as that needed to raise the temperature from 14°C to 15°C and so on. There are, however, two limitations to an interval scale. The first is that zero on the scale does not mean absence of the quantity being measured: 0°C does not mean complete absence of heat. Secondly, we cannot make simple ratio statements such as a day of 40°C is twice as hot as a day of 20°C.
- A **ratio** scale has all the properties of an interval scale but has the additional properties that zero means complete absence of what is being measured, and ratio statements, such as ‘half as many’ and ‘fifty times as much,’ can be made. Ratio scales are those with which you are probably most familiar. When the number of brothers are counted or a person’s hand span is measured in cm, you are using a ratio scale. On a ratio scale, zero means complete absence of the quantity, for example, zero brothers. With a true zero point, you can make true ratio statements; a person with 4 brothers has twice as many brothers as someone who has 2 brothers.

For statistical purposes it is generally not necessary to distinguish between interval and ratio scales as both give numbers on which we can perform the same statistical analyses.

Types of variables

Categorical variables

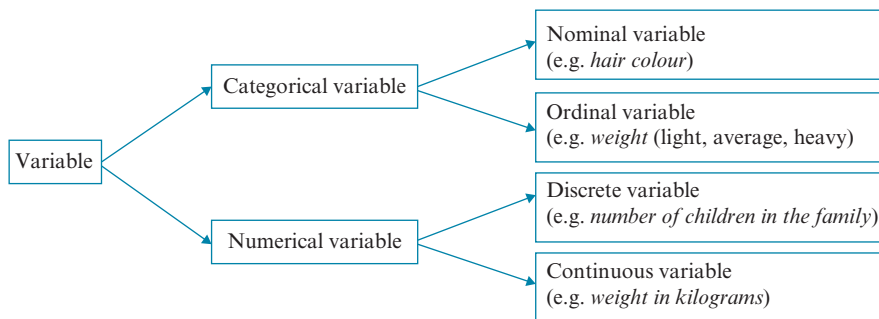
Variables that generate categorical data are called **categorical variables**. We can further separate categorical variables into **nominal** or **ordinal** variables. For example, *gender* is a nominal variable, while *coffee size* is an ordinal variable.

Numerical variables

Variables that generate numerical data are called **numerical variables**. We can further separate numerical variables into discrete or continuous variables. For example, *number of brothers* is a **discrete variable**, while *hand span* is a **continuous variable**.

Numerical or categorical?

The relationship between categorical variables (nominal or ordinal) and numerical variables (discrete or continuous) is displayed in the following diagram:



Example 1 Categorical and numerical variables

Classify the following variables as categorical or numerical.

- a** Students choose their favourite pet from ‘dog’, ‘cat’, ‘bird’ or ‘other’.
- b** The time, in seconds, taken to solve a puzzle is recorded.

Solution

- a** **Categorical**, as the values of the variable are categories of pet.
- b** **Numerical**, as the data takes values which represent the amount of time taken.



**Example 2** Nominal and ordinal variables

Classify the following variables as nominal or ordinal.

- a** A group of people record their level of happiness as ‘very happy’, ‘happy’, ‘not too happy’ or ‘very unhappy’.
- b** Students select their favourite country to visit.

Solution

- a Ordinal**, as the data takes values which represent the person’s level of happiness, and there is an order to the categories.
- b Nominal**, as the data takes values which are names of countries.

**Example 3** Discrete and continuous variables

Classify the following numerical variables as discrete or continuous.

- a** The number of children in the family is recorded for all the students in a school.
- b** The birth weight of babies, measured in grams, is recorded at a hospital.

Solution

- a Discrete**, as the number of children will only take whole number values.
- b Continuous**, as the data can take any value, limited only by the accuracy to which the weight can be measured.

**Example 4** Classifying variables

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

- a** The number of students in each of 10 classes is counted.
- b** The time taken for 20 mice to each complete a maze is recorded in seconds.
- c** Diners at a restaurant were asked to rate how they felt about their meal: 1 = Very satisfied, 2 = Satisfied, 3 = Indifferent, 4 = Dissatisfied, 5 = Very dissatisfied.
- d** Students choose a colour from a list: 1 = Blue, 2 = Green, 3 = Red, 4 = Yellow.
- e** Students’ heights were classified as ‘less than 160 cm’, ‘160 cm - 180 cm’ or ‘more than 180 cm’.

Solution

- a Discrete**, as the number of students will only take whole number values.
- b Continuous**, as the data can take any value, limited only by the accuracy to which the time can be measured.
- c Ordinal**, as the numbers in this data do not represent quantities, they represent each diner’s level of approval of the meal.
- d Nominal**, as the numbers do not represent quantities, they represent colours.
- e Ordinal**, as each student’s height is recorded into 3 categories which can be ordered.

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.

Note that it is quite common for percentages to add to 99.9% or 100.1%. This is due to the rounding of each of the individual percentages in the table and is not a concern.



Example 5 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.

Explanation

- 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.

Solution

<i>Lunch choice</i>	Frequency	
	Number	%
Sandwich	7	23.3
Salad	10	33.3
Pie	13	43.3
Total	30	99.9

Now try this 5 Constructing frequency table for categorical data (Example 5)

Twenty-five students were asked how they usually travelled to school. They answered as follows:

walk, car, car, other, car, car, walk, car, bus, bus, car, bus, walk, walk, bus, bus, bus, car, car, car, car, car, bus, bus

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

Hint 1 Determine the possible values of the variable *Travel mode* - there are four.

Hint 2 Check that the frequencies add to 25.

Hint 3 Check that the percentage frequencies add to 100%.

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 6 Constructing bar and percentage bar charts from a frequency table

Use the frequency table for *Lunch choice* from Example 5 to construct:

a a bar chart

b a percentage bar chart.

Explanation

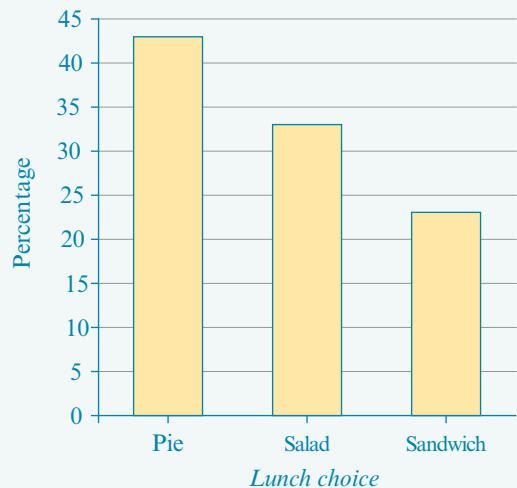
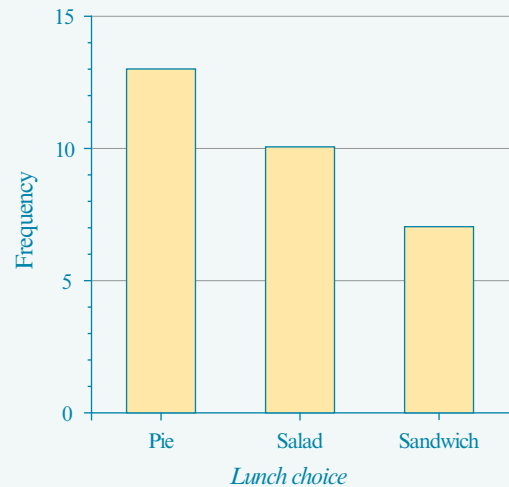
a 1 Label the horizontal axis with the variable name, *Lunch choice*. Mark the scale off into three equal intervals and label them 'Pie', 'Salad' and 'Sandwich'.

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 intervals of 5.

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.

b 1 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%.

2 Mark the vertical scale in intervals of 5%. The height of each bar is equal to the percentage frequency.

Solution


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

Now try this 6**Constructing bar and percentage bar charts from a frequency table (Example 6)**

Use the frequency table for *Travel mode* from Now Try This (Example 5) to construct:

- a** a bar chart
- b** a percentage bar chart.

Hint 1 The horizontal axis for both charts is labelled with the values of the variable *Travel mode*.

Hint 2 The vertical axis of the bar chart is labelled Frequency. The scale should start at 0 and extend slightly more than the maximum frequency.

Hint 3 The vertical axis of the percentage bar chart is labelled Percentage. The scale should start at 0 and extend slightly more than the maximum percentage frequency.

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 5, 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?’

Section Summary

- ▶ Data can be classified as **categorical** (nominal or ordinal), or **numerical** (discrete or continuous).
 - ▶ **Nominal data** takes values which are simply the **names** of categories.
 - ▶ **Ordinal data** takes values which both **name** and **order** categories.
 - ▶ **Discrete** data can only take particular numerical values, often whole numbers.
 - ▶ **Continuous** data can take any numerical value within a specified range.
- ▶ **Categorical** data can be summarised in a **frequency table** or a **percentage frequency table**.
- ▶ A frequency table can be displayed in a **bar chart**.
- ▶ A percentage frequency table can be displayed in a **percentage bar chart**.
- ▶ The value of a categorical variable with the highest frequency is called the **mode**.



Exercise 2A

Building understanding

Example 1

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



Example 2

- 2 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’?

Example 3

- 3 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.

Example 4

- 4** 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.
 - d** The annual salaries of the employees of a certain company were recorded as 'less than \$60 000', '\$60 000 - \$90 000', 'more than \$90 000'.

Developing understanding**Example 5**

- 5** A group of people were asked to identify their gender as F = female or M = Male. The following data was collected:

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

- a** Is the variable *gender* nominal or ordinal?
 - b** Construct a frequency table for the data including frequency and percentage frequency.
- 6** A group of 18-year-old males were asked to give their shoe size. The following data was collected:
- 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.

Example 6

- 7** The table shows the frequency distribution of the favourite type of fast food (*Food type*) of a group of students.

- 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 this group of students?
- f** Construct a frequency bar chart.

<i>Food type</i>	Frequency	
	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

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

- 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 frequency bar chart.

<i>Attitude to capital punishment</i>	Frequency	
	Number	%
Strongly agree	21	8.2
Agree	11	4.3
Don't know	42	
Disagree		
Strongly disagree	129	50.4
Total	256	100.0

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

- 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 (%).

<i>Type of book</i>	Frequency	
	Number	%
Children	53	22.8
Fiction	89	
Cooking	42	18.1
Travel	15	
Other	33	14.2
Total	232	

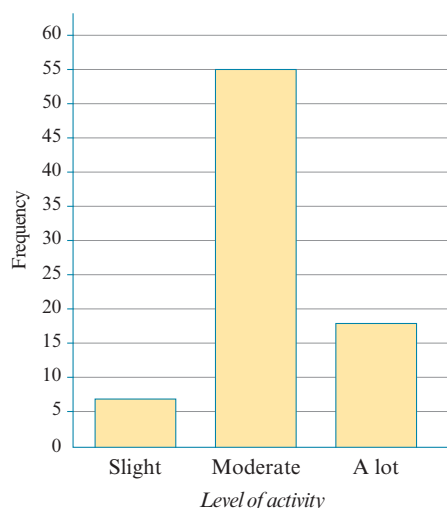
- 10** The members of a club were classified according to the following age groups.

- a** How many people are in the club?
b Is the variable *Age group* nominal, ordinal or numerical?
c What percentage of people in the club were aged 35-44 years?
d What is the modal age category?
e Construct a bar chart of the percentage frequencies (%).

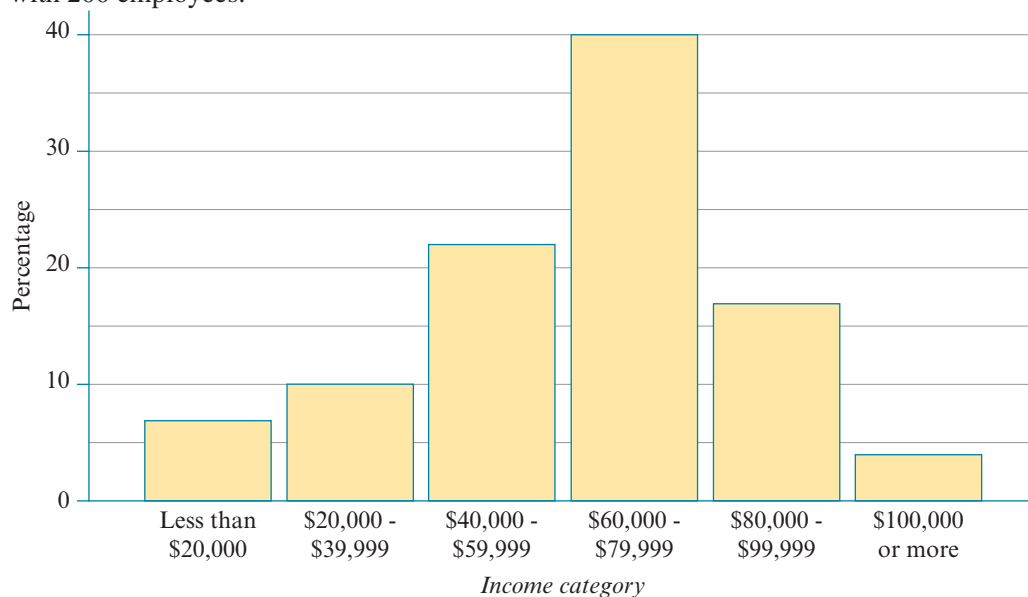
<i>Age group</i>	Frequency	
	Number	%
Under 18	84	42
18 - 24	26	13
25 - 34	46	23
35 - 44	24	12
45 - 54	8	4
55 or over	12	6
Total	200	100

Testing understanding

- 11** In a survey, people were asked to select their level of activity as ‘slight’, ‘moderate’ or ‘a lot’. Their responses are summarised in the following bar chart.



- Give the name of the variable, and classify it as nominal, ordinal or numerical.
 - How many people responded to this question in the survey?
 - How many people responded that they exercise ‘a lot’?
 - What is the modal category for the variable, and what percentage of people chose that response?
- 12** The following percentage bar chart shows the incomes of people in a company with 200 employees.



- Is the variable *Income category* nominal, ordinal or numerical?
- How many employees earned salaries in the modal income category?

2B Interpreting and describing frequency tables and bar charts

Learning intentions

- ▶ To be able to explain the features of a categorical variable by interpreting frequency tables and bar charts.
- ▶ To be able to write a report which communicates your findings.

As part of this topic, 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 (more than 3), 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.



Example 7 Describing the distribution of a categorical variable from a frequency table

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

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

<i>Lunch choice</i>	<i>Frequency</i>
Sandwich	7
Salad	10
Pie	13
Total	30

Continued

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 a pie, chosen by 13 of the children. Ten children chose a salad. The least popular option was a sandwich, chosen by only 7 of the children.

Now try this 7**Describing the distribution of a categorical variable from a frequency table (Example 7)**

A group of 25 kindergarten children were asked to choose an activity from painting, story time and playdough. Their choices are summarised in the frequency table opposite.

<i>Activity choice</i>	Frequency
Painting	8
Story time	3
Playdough	14
Total	25

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

Hint 1 Make sure you mention the modal value as ‘the most popular choice’ and the lowest frequency as ‘the least popular choice’.

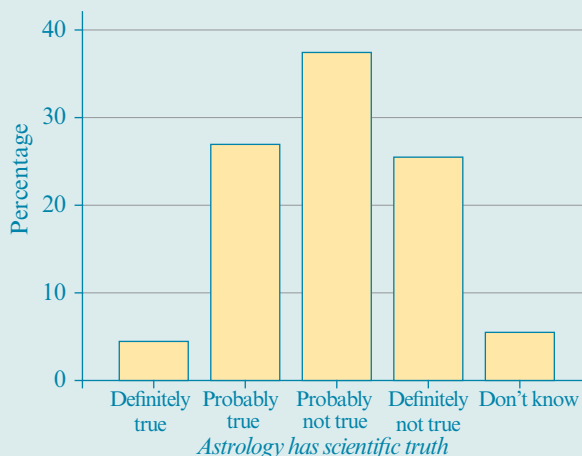
Hint 2 Write as if you are explaining the children’s choices to a friend.

**Example 8****Describing the distribution of a categorical variable from a frequency table and bar chart**

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 (in a definite order because the data are ordinal).

<i>Astrology has scientific truth</i>	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 using the frequency table and bar chart.

Solution**Report**

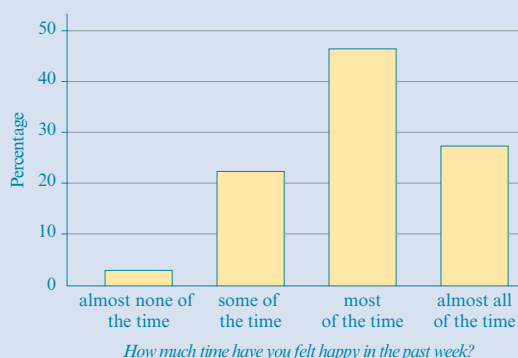
A sample of two hundred people were asked to respond to the statement ‘Astrology has scientific truth’.

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% thought that the statement was definitely true.

Now try this 8**Describing the distribution of a categorical variable from a frequency table and bar chart (Example 8)**

A sample of 66 people were asked to respond to the question ‘How much time have you felt happy in the past week’ by selecting one of the options ‘almost none of the time’, ‘some of the time’, ‘most of the time’ or ‘almost all of the time’. The data are summarised in the frequency table and bar on the following page. Write a report summarising the findings of this investigation, quoting appropriate percentages to support your conclusions.

How much happy time?	Frequency	
	Number	%
almost none	2	3.0
some	15	22.7
most	31	47.0
almost all	18	27.3
Total	66	100.0



Hint 1 Compare the percentage frequencies associated with each value of the variable from highest to lowest.

Hint 2 Make sure you use comparative terms such as ‘more’ and ‘less’ - don’t just state the percentages.

Hint 3 Write as if you are explaining the responses to this question to another person.

Section Summary

- ▶ Examination of frequency tables and bar charts can help us to understand the distribution of a categorical variable.
- ▶ Important features such as the values of the variable with the highest and lowest frequencies should be included in a written report.
- ▶ Be sure to always include the total number of data values.

Exercise 2B

Building understanding

Example 7

- 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.

<i>Fast food type</i>	<i>%</i>
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.

<i>Capital punishment</i>	<i>%</i>
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.

Developing understanding

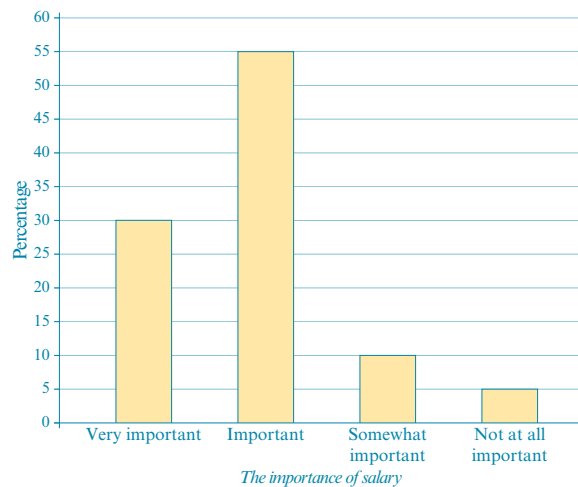
- 3 A group of 200 students were asked how they prefer to spend their leisure time. The results are summarised in the frequency table opposite.

Use the information in the table to write a brief report on the results of this investigation.

<i>Leisure activity</i>	<i>%</i>
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 8

- 4 A group of 600 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 bar chart.



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

Testing understanding

- 5 Ask your class (or a convenient group of people) to respond to the question ‘How concerned are you about climate change?’ by selecting one of a range of responses (ensure that you have about four or five options).
- a Summarise the responses in a frequency table with percentages and with a bar chart.
 - b Write a report based on your findings.

2C Displaying and describing numerical data

Learning intentions

- ▶ To be able to construct frequency tables for discrete numerical data.
- ▶ To be able to construct frequency tables for grouped numerical data (discrete and continuous).
- ▶ To be able to construct a histogram from frequency tables for numerical data.
- ▶ To be able to construct a histogram from numerical data using a CAS calculator.

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

Discrete data



Example 9 Constructing a frequency table for discrete numerical data, taking a small number of values

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 table for these data showing both frequency and percentage frequency.

Explanation

- 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.
- 6 Total the percentages and record.

Solution

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

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

Now try this 9 Constructing a frequency table for discrete numerical data, taking a small number of values (Example 9)

The number of faulty widgets produced each hour over a 24-hour period by a certain machine are as follows:

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

- Hint 1** Determine the minimum and maximum values which the variable *faulty widgets* can take from the data set.
- Hint 2** When you have counted the frequencies for each value, check that they add to 24.
- Hint 3** Check that the percentage frequencies add to 100%.

Grouping data

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

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

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

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

- A division which results in about 5 to 15 groups is preferred.
- 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 10** Constructing a grouped 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 grouped frequency table of these data showing both frequency (count) and percentage frequency.

Explanation

- 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 and 4.
- Set up the table as shown.
- Count the data values in each interval to complete the number column.
- Convert the frequencies into percentages and record in the per cent (%) column.
For example, for the interval 5–9: % frequency = $\frac{3}{20} \times 100 = 15\%$.
- Total the percentages and record.

Solution

<i>Cups of coffee</i>	Frequency	
	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

Now try this 10 Constructing a grouped frequency table for a discrete numerical variable (Example 10)

A group of thirty people were each asked the number of times they dined in a restaurant in the last 3 months, with the following results:

1 6 9 12 13 29 0 5 44 52
7 0 10 17 3 6 0 53 23 9
5 2 18 21 6 9 2 63 58 0

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

Hint 1 Group the data in intervals of width 10.

Hint 2 When you have counted the frequencies for each group, check that they add to 30.

Hint 3 Check that the percentage frequencies add to 100%.

Grouped continuous data

**Example 11** 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 and a percentage frequency table for these data.

Explanation

- 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.
- 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.
- 6** Total the percentages and record.

Solution

Height	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

For example, for the interval 175.0–179.9:

$$\% \text{ frequency} = \frac{5}{41} \times 100 = 12.2\%.$$

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

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 12 Constructing a histogram for ungrouped discrete data

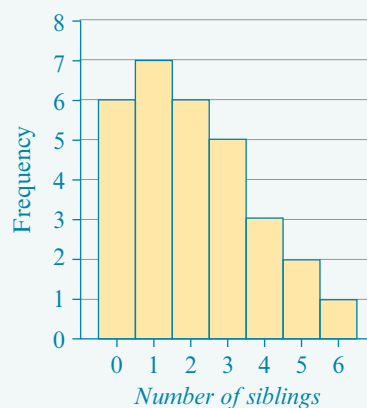
Construct a histogram for the data in the frequency table.

<i>Number of Siblings</i>	Frequency
0	6
1	7
2	6
3	5
4	3
5	2
6	1
Total	30

Explanation

- 1** Label the horizontal axis with the variable name *Number of siblings*. Mark in the scale in units that include 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.

Solution



- 3** For each value of 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.

Now try this 12**Constructing a histogram for ungrouped discrete data (Example 12)**

Use the frequency table for *Faulty widgets* from Now Try This (Example 9) to construct a histogram for the data.

Hint 1 The horizontal axis is labelled *Number of faulty widgets*.

Hint 2 The vertical axis of the histogram is labelled Frequency. The scale should start at 0 and extend slightly more than the maximum frequency.

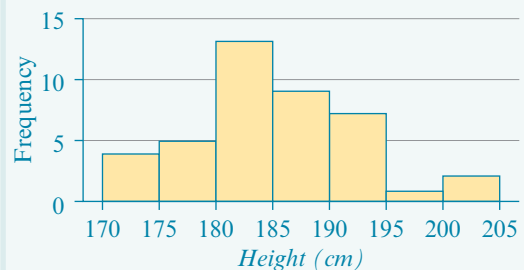
**Example 13****Constructing a histogram for continuous data**

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

Explanation

- Label the horizontal axis with the variable name *Height (cm)*. Mark in the scale using the beginning of each interval as the scale points; that is, 170, 175, ...
- 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.
- 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.

Solution

Now try this 13 Constructing a histogram for continuous data (Example 13)

The number of hours per week spent on email by a group of 150 people are summarised in this frequency table. Use it to construct a histogram for the data.

<i>Hours on email</i>	Frequency
0.0–4.9	47
5.0–9.9	52
10.0–14.9	21
15.0–19.9	9
20.0–24.9	7
25.0–29.9	5
30.0–34.9	5
35.0–39.9	2
40.0–44.9	0
45.0–49.9	2
Total	150

Hint 1 Label the horizontal axis with the variable name *Hours on email*. Mark in the scale using the beginning of each interval as the scale points; that is, 0, 5, 10, ...

Hint 2 Label the vertical axis 'Frequency'. The scale should start at 0 and extend slightly more than the maximum frequency.

Hint 3 There should be no gaps between the columns in the histogram.

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.

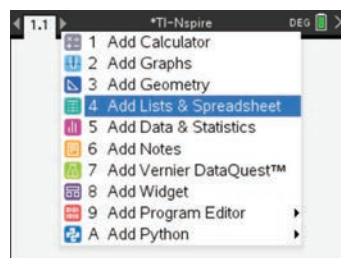
How to construct a histogram using the TI-Nspire CAS

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 **Ctrl** + **on** and select **New** (or use **Ctrl** + **N**). If prompted to save an existing document, move the cursor to **No** and press **enter**.
- 2 Select **Add Lists & Spreadsheet**.
Enter the data into a list named *marks*.



- a Move the cursor to the name cell of column A (or any other column) and type in *marks* as the list variable. Press **enter**.
- b Move the cursor down to row 1, type in the first data value and press **enter**. Continue until all the data has been entered. Press **enter** after each entry.

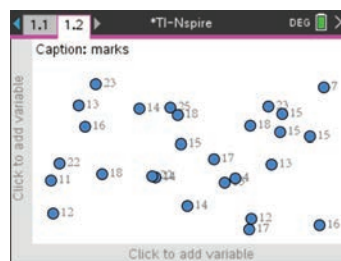
	A marks	B	C	D
10	12			
11	15			
12	13			
13	16			
14	14			

3 Statistical graphing is done through the **Data & Statistics** application.


Press **ctrl** + **doc** (or alternatively press **ctrl** + **I**) and select **Add Data & Statistics** (or

press **on**, arrow to , and press **enter**).

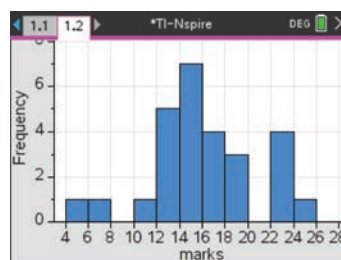
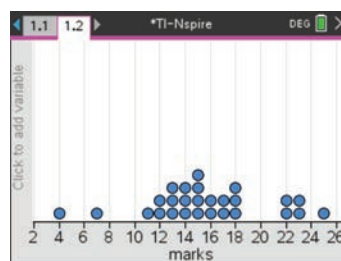
Note: A random display of dots will appear – this is to indicate that data are available for plotting. It is not a statistical plot.




- 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.

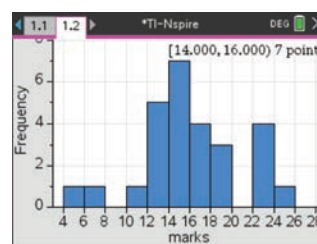
- b A dot plot is displayed as the default plot. To change the plot to a histogram, press **menu** > **Plot Type** > **Histogram** and then press **enter** or 'click' (press ).

Your screen should now look like that shown opposite. This histogram has a column (or bin) width of 2 and a starting point of 4.



4 Data analysis

- a Move the 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.

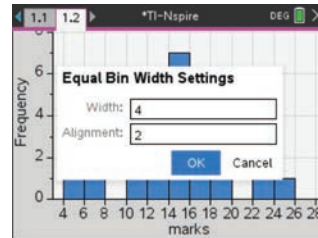
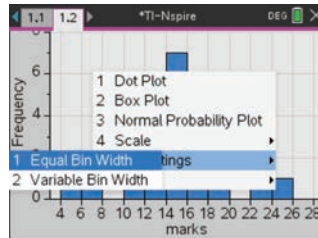
5 Change the histogram column (bin) width to 4 and the starting point to 2.

- a** Press **ctrl** + **menu** to access the context menu as shown (below left).

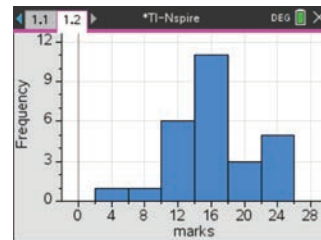
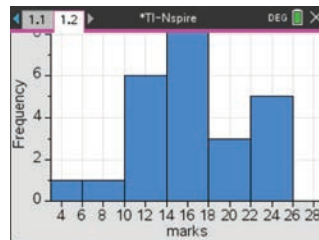
Hint: Pressing **ctrl** + **menu** with the cursor on the histogram gives you access to a context menu that enables you to do things that relate only to histograms.

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

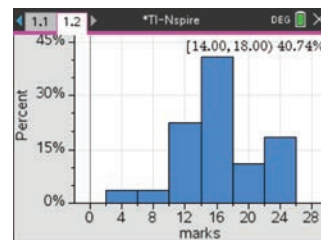
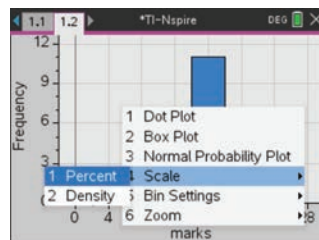
- c** In the settings menu (below right) change the **Width** to **4** and the **Starting Point (Alignment)** to **2** as shown. Press **enter**.



- d** A new histogram is displayed with a 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

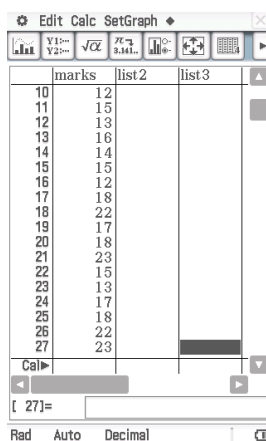
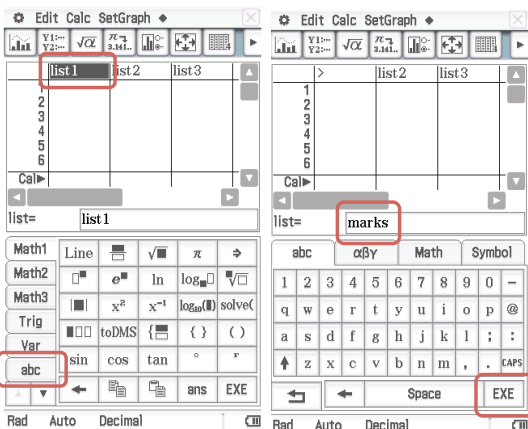
Steps

- From the application menu screen, locate the **Statistics** application.
Tap  to open.


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

- Enter the data into a list named *marks*.
 - Tap on the column heading list 1.
 - Press **Keyboard** and tap **abc**.
 - Type *marks* and press **EXE**.
 - 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.




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  to confirm your selections.

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**.

5 The screen is split in two.

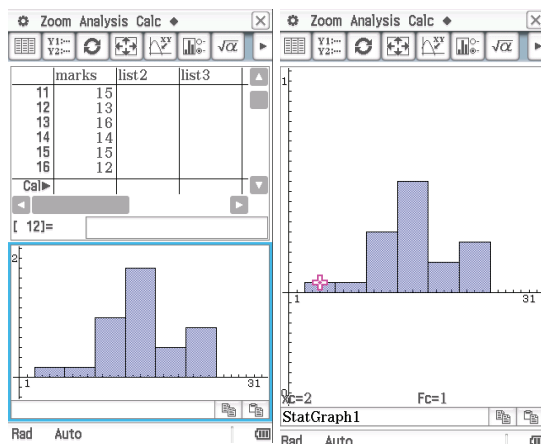
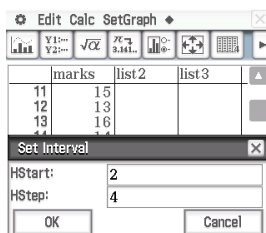
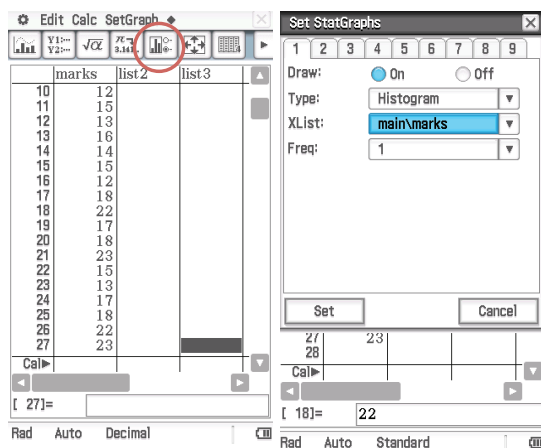
Tapping  from the icon panel will allow the graph to fill the entire screen.

Tap  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 ($F_c = 1$).

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



Section Summary

- ▶ Numerical data (both discrete and continuous) can be summarised in frequency tables (counts or percentages).
- ▶ When the data is **discrete, but can take many values**, the data should be **grouped** to form the frequency table.
- ▶ When the data is **continuous**, the data should be **grouped** to form the frequency table.
- ▶ A **histogram** is a display of a frequency table of numerical data.
- ▶ A CAS calculator can be used to construct a histogram.

Exercise 2C

Building understanding

Example 9

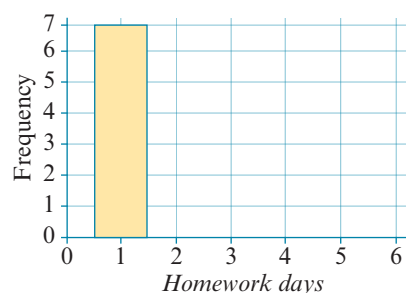
- 1 A group of 20 students were surveyed about the number of days they completed homework over a one-week period. The data is as shown.

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

- a Use the data to complete the frequency table.

Homework days	Frequency
1	7
2	
3	
4	
5	
Total	20

- b Use the information in the frequency table to complete the histogram shown.



Example 10

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

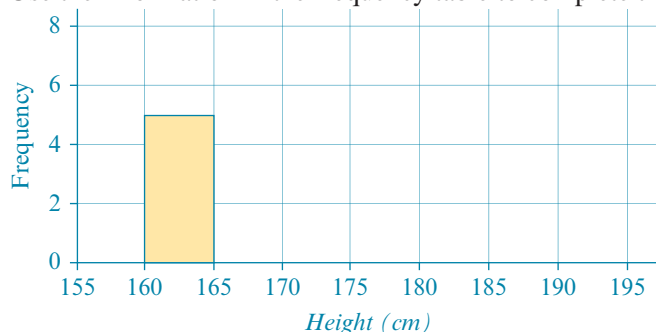
Example 12

- a Use the data to complete the grouped frequency table.

188 175 176 161 183
 169 171 176 165 166
 162 170 174 168 178
 169 180 173 163 179
 163 170 163 175 177

Height (cm)	Frequency
160–164	5
165–169	
170–174	
175–179	
180–184	
185–190	
Total	25

- b** Use the information in the frequency table to complete the histogram shown.



Developing understanding

- 3** 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 11

- 4** 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.

Example 13

- 5** 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.

- 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?

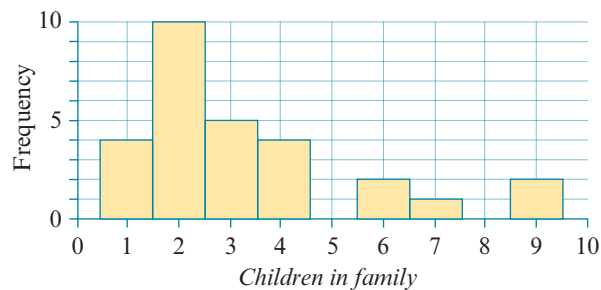
- b** What percentage of students drew a line with a length:

- i** from 31.0 to 31.9 cm?
- ii** of less than 31 cm?
- iii** of 30 cm or more?

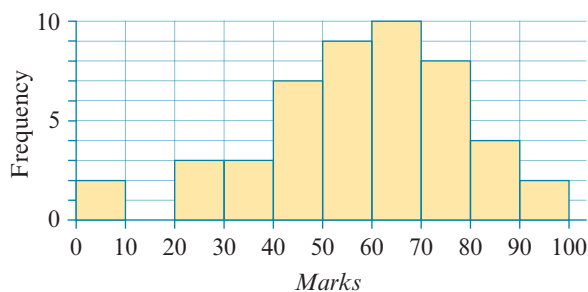
<i>Length of line (cm)</i>	Frequency	
	Number	%
28.0–28.9	1	3.6
29.0–29.9	2	7.1
30.0–30.9	8	28.6
31.0–31.9	9	32.1
32.0–32.9	7	25.0
33.0–33.9	1	3.6
Total	28	100.0

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

- 6** 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?
- 7** 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?
e Of this group of students, what percentage did not pass the quiz? Round your answer to the nearest whole number.

- 8** 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 a column width of 10 and a starting point of 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?
- 9** 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 a column width of 2 and a starting point of 0. Name the variable *Max temp*.
- 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?

Testing understanding

- 10** The number of mistakes made on a test by each of 30 students is as follows.

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

- a** Organise the data into a frequency table.
- b** Construct a histogram of the data from the frequency table.
- c** What is the mode?
- d** A student who makes five or more errors is required to re-sit the test. What percentage of students would be required to re-sit?

- 11** The following data gives the waiting time, in minutes, for a group of 30 people who attended an emergency department of a hospital.

36	76	14	26	11	90	32	2	16	125
12	45	59	75	17	5	18	22	45	28
15	23	9	53	17	13	31	78	12	51

- Use the data to construct a frequency table.
- Construct a histogram of the data from the frequency table.
- What is the modal value for the variable, and what percentage of people chose that response?

2D Characteristics of distributions, dot plots and stem plots

Learning intentions

- ▶ To be able to identify the key characteristics of a data distribution.
- ▶ To be able to construct a dot plot and a stem plot for numerical data.

Characteristics of a distribution

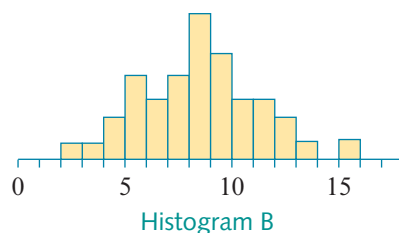
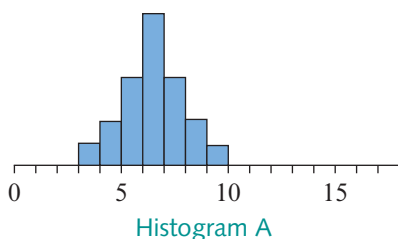
Distributions of numerical data are characterised by their shape and special features such as location (also referred to as the ‘centre’) and spread.

Shape of a distribution

Symmetry

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

Histogram A below is exactly symmetric, while Histogram B shows a distribution that is approximately symmetric. In practice it is rare to find a histogram which is exactly symmetric, and approximate symmetry is enough for us to classify a histogram as symmetric.

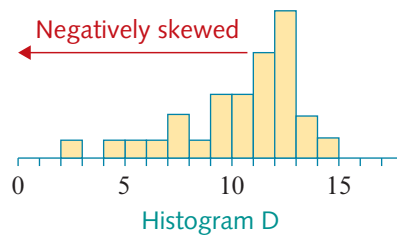
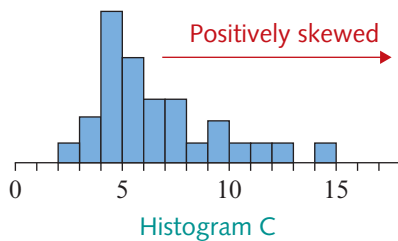


Positive and negative skew

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.



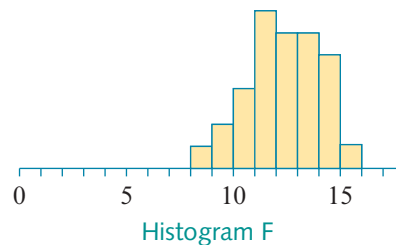
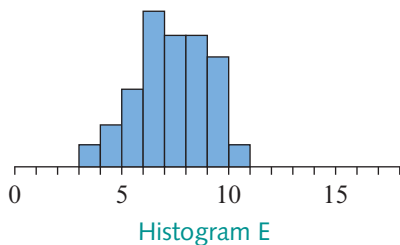
Knowing whether a distribution is skewed or symmetric is important, as this gives considerable information concerning the choice of appropriate summary statistics, as will be seen in the next section.

Centre

Comparing centre

Two distributions are said to differ in **centre** 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 is moved horizontally several units to the right, indicating that these distributions differ in location.

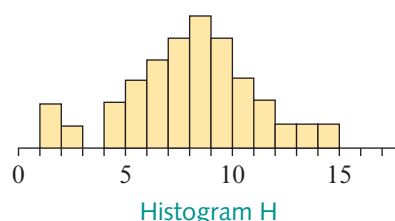
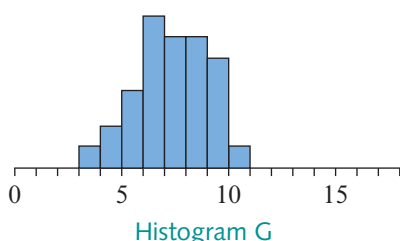


Spread

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.



As we have seen here, a histogram enables us to identify and compare the characteristics of a data distribution (shape, centre and spread). These key features can also be seen in two other plots which you are already familiar with, the dot plot and the stem and leaf plot.

Dot plots

The simplest display of numerical data, and an alternative to a frequency histogram, is the **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 display fairly small data sets where the data takes a limited number of values.



Example 14 Constructing a dot plot

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

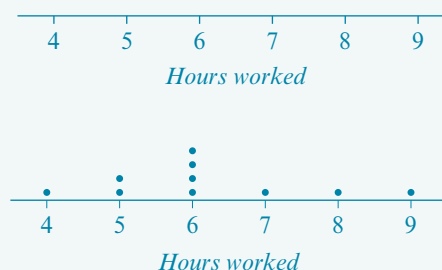
6 9 5 8 6 4 6 7 6 5

Construct a dot plot of these data.

Explanation

- 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.

Solution



In the same way that the shape of a distribution can be identified from a histogram, it can also be identified from the dot plot (if there are enough data values). When there is a longer tail on the dot plot in the positive direction, then the distribution would be positively skewed, while if the tail on the dot plot is in the negative direction, then the distribution would be negatively skewed. The dot plot in Example 14 would be considered to be approximately symmetric, although when there are so few data values it is difficult to comment on shape with any certainty.

Now try this 14 Constructing a dot plot (Example 14)

The number of pages written for an assignment by a group of 12 students is as follows:

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

Construct a dot plot of these data.

Hint 1 Construct a number line spanning the minimum and maximum data values.

Hint 2 Ensure the vertical distance between dots is consistent.

Stem plots

The **stem plot** or **stem and leaf plot** is another very useful plot for displaying small numerical data sets.

Stem plot

A stem plot is a display where each data value is split into a **stem** (usually the leading digit or digits) and a **leaf** (usually the last digit). For example, 45 is split into 4 (stem) and 5 (leaf).

The stem values are listed vertically, and the leaf values are listed horizontally next to their stem. The stem is usually separated from the leaves by a vertical line.

**Example 15** 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 18 60 42 37 28
28 17 19 52 55 27 5 7 19 45 19 25

Display the data in the form of an ordered stem plot, and comment on the shape of the distribution.

Explanation

- 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 (smallest to largest) 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.

Solution

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

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 5 7
1 | 5 9 8 7 9 9 9
2 | 4 5 4 8 8 7 5
3 | 0 3 7
4 | 2 5
5 | 2 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 key: 1 | 5 = 15 marks

```

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

```

- 4** Looking at the stem plot, we can see that the tail of the distribution is longer in the direction of the increasing test scores (if this isn't clear then turn the stem plot on its side), so we can say this distribution is positively skewed.

Note that the stem is defined as the leading digit or digits and the leaf as the final digit. That is why it is always necessary to include a key, so that we know how to interpret the stem and the leaves.

Now try this 15 Constructing a stem plot (Example 15)

The weights of each of 22 pumpkins (in kg) grown in Essie's garden are as follows:

1.9 4.2 2.4 3.0 2.9 3.4 4.4 3.3 4.1 6.0 5.5
 2.5 2.8 3.8 3.7 2.8 4.5 2.0 6.8 7.0 4.3 4.5

Display the data in the form of an ordered stem plot.

Hint 1 Choose values for the stem which span the minimum and maximum data values.

Hint 2 Make sure you include a key.

Choosing between plots

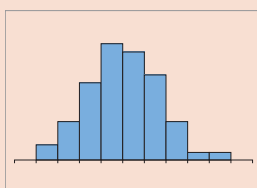
We now have three different plots that can be used to display numerical data: the histogram, the dot plot and the stem and leaf plot. They all allow us to make judgements concerning the important features of the distribution of the data, so how would we decide which one to use?

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

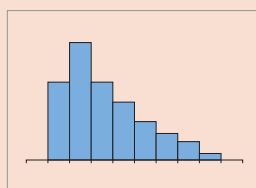
Plot	Used best when	How usually constructed
Dot plot	small data sets (say $n < 30$) discrete data	by hand or with technology
Stem plot	small data sets (say $n < 50$)	by hand
Histogram	large data sets (say $n > 30$)	with technology

Section Summary

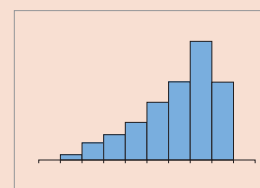
- ▶ **Numerical data distributions** are characterised by shape (symmetric or skewed), location and spread.
- ▶ The following shapes are commonly seen in data distributions.



symmetric



positively skewed



negatively skewed

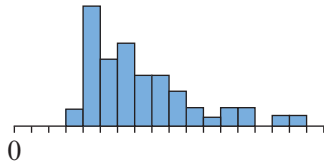
- ▶ A **dot plot** is an appropriate display for a data distribution when there are a small number of data values.
- ▶ A **stem plot** is an appropriate display for a data distribution when there are a small to medium number of data values.
- ▶ A **histogram** is an appropriate display for a data distribution when there are a medium to large number of data values, and technology is used.
- ▶ The characteristics of a data distribution (shape, centre and spread) can be visually identified from all of these plots.

Exercise 2D

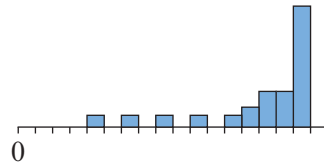
Building understanding

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

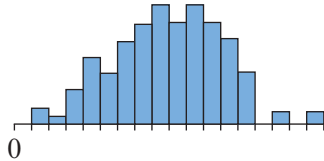
a



b

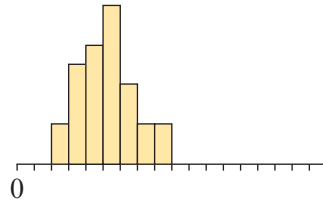
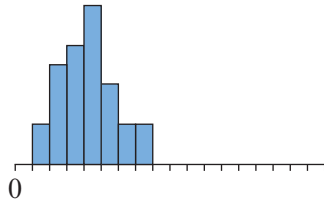


c

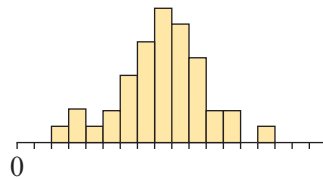
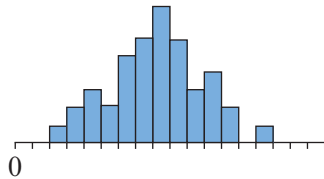


- 2 Do the following pairs of distributions differ in spread, centre, both or neither? Assume that each pair of histograms is drawn on the same scale.

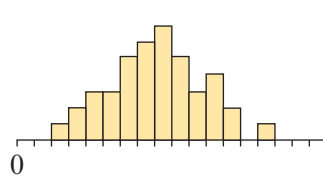
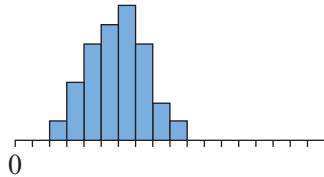
a



b



c



Developing understanding

Example 14

- 3 The number of children in each of 15 families is as follows:

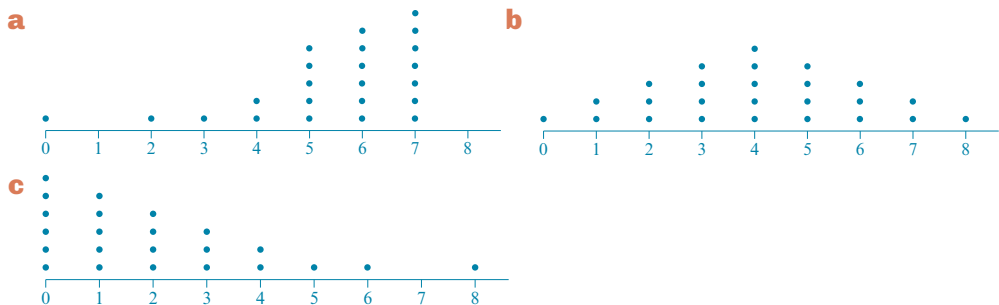
0 7 2 2 2 4 1 3 3 2 2 2 0 0 1

- a Construct a dot plot of the number of children.
b What is the mode of this distribution?

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



- 6** The ages of each member of a football team are as follows:

22 20 20 21 21 22 24 25 25 24 26
 22 19 28 24 25 30 21 27 24 25 26

- a** Construct a dot plot of the ages of the players.
b What is the mode of this distribution?
c What is the shape of the distribution of player ages?
d What percentage of players are younger than 25?
- 7** In a study of the service offered at her cafe, 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 cafe seem to be?

Example 15

- 8** 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 What percentage of students obtained a mark of 50 or more?
c What was the lowest mark?

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

a key: 3|4 represents 34

```

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

b key: 3|1 represents 31

```

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

c key: 1|0 represents 10

```

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

- 10** 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)

key: 1 | 2 means 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
  
```

- 11** 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?

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

```

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?

- 13** 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?

Testing understanding

- 14 a** The minimum daily temperature over a two-week period in a certain town was recorded as follows:

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

Construct a dot plot of the data.

- b** The maximum daily temperature over the same two-week period in that town was also recorded as follows:

12 14 13 13 13 15 15 17 16 13 13 12 13 13

Construct a dot plot of the data, using the same scale for the axes as in part **a**.

- c** How do the two distributions compare in terms of centre and spread?
- 15** A researcher determined the percentage body fat for 30 adult males, as follows:
- 5 12 9 17 19 5 7 21 30 24 9 21 10 20 17
18 28 11 14 17 13 4 20 17 25 29 17 26 10 13
- a** Construct a stem plot of the data.
 - b** After the researcher had finished collecting the data, she noted that there had been an error in the data collection which could be fixed by subtracting 2 from each data value. How do you think the distribution of the corrected data would compare to that shown in part **a** in terms of centre and spread?
- 16** State whether you would predict the data distribution of the following variables to be symmetric, positively skewed or negatively skewed. Give a reason for your choice.
- a** The height in cm of 16-year-old boys.
 - b** The purchase price, in dollars, of a house in Melbourne.
 - c** The gestation period for a human baby in weeks.

2E Measures of centre

Learning intentions

- ▶ To be able to understand the mean and the median as measures of centre.
- ▶ To be able to know whether to use the median or mean as a measure of centre for a particular distribution.

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.

In this section we will consider some summary statistics which are 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 commonly referred to as 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, Σ , as a shorthand way of writing ‘sum of’
- a lower case x , to represent a data value
- a lower case x with a bar, (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: $\bar{x} = \frac{\Sigma x}{n}$


Example 16 Calculating the mean

The following data set shows the number of premierships won by each of the current AFL teams until the end of 2021. Find the mean of the number of premierships won. Round your answer to one decimal place.

Team	Premierships
Carlton	16
Essendon	16
Collingwood	15
Melbourne	13
Hawthorn	13
Richmond	13
Brisbane Lions	11
Geelong	9
Sydney	5

Team	Premierships
Kangaroos	4
West Coast	4
Adelaide	2
Western Bulldogs	2
Port Adelaide	1
St Kilda	1
Fremantle	0
Gold Coast	0
GWS	0

Explanation

- 1 Write down the formula and the value of n .
- 2 Substitute into the formula and evaluate.
- 3 We do not expect the mean to be a whole number, so give your answer to one decimal place.

Solution

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} & n &= 18 \\ \bar{x} &= \frac{16 + 16 + 15 + \dots + 1 + 1 + 0 + 0 + 0}{18} \\ &= \frac{125}{18} \\ &= 6.9\end{aligned}$$

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 the median 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; since 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 values (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 fewer times.

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

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 $\left(\frac{n+1}{2}\right)$ 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 $\frac{n}{2}$ th and the $\left(\frac{n}{2} + 1\right)$ th observations are added together and divided by 2.



**Example 17** Determining the median when n is odd

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

Age (years)	key: 1 2 means 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

Explanation

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

- 1 Determine the number of observations.
- 2 Since there are an odd number of values, the median is located at the $\frac{n+1}{2}$ th position.

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.

Solution

$$n = 23$$

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

Thus the median age is 41 years.

**Example 18** Determining the median when n is even

Find the median age for a group of people whose ages are displayed in this ordered stem plot.

Age (years)	1 2 represents 12 years
0	5 9
2	1 3 5 8
3	0 0 4 9 9
4	0 4 5 8
5	3 7
6	3

Explanation

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

- 1 Determine the number of observations.
- 2 Since there are an even number of observations, to find the median the $\frac{n}{2}$ th and the $(\frac{n}{2} + 1)$ th observations are added together and divided by 2.

Solution

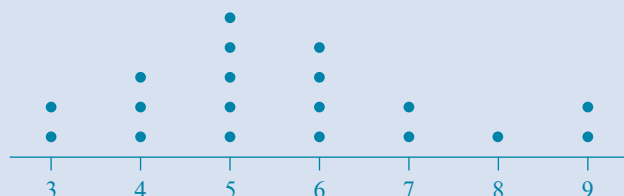
$$n = 18$$

Median is the average of the values in the $\frac{18}{2} = 9$ th and $\frac{18}{2} + 1 = 10$ th positions.

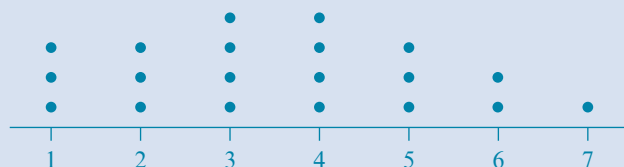
Thus the median age is $\frac{34 + 39}{2} = 36.5$ years.

Now try this 18 Determining the median from a dot plot (Examples 17 and 18)

1 Find the median of the data displayed in the dotplot shown.



2 Find the median of the data displayed in the dotplot shown.



Hint 1 Determine the number of data values in the dot plot.

Hint 2 Locate the median using the appropriate formula for an odd or even number of data values.

Comparing the mean and median

In Example 16 we found that the mean number of premierships won by the 18 AFL clubs was $\bar{x} = 6.9$. We also 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	2	2	4	4
0	5	9							
1	1	3	3	3					
1	5	6	6						

From the stem and leaf plot it can be seen that the distribution of premierships won 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 mean is pulled towards the tail of the distribution and the extreme values, giving us a value for the mean which may be far from the centre. Since the median is not so affected by unusual observations and always gives the middle value, then the median is the preferred measure of centre for a skewed distribution or one with outliers. (Outliers are really large or really small numbers compared to the rest of the data.) When the distribution is symmetric, and there are no outliers, then either measure is appropriate, although we tend to prefer the mean as it is easier to calculate and more familiar to most people.

Section Summary

- ▶ **Summary statistics** are numbers calculated from a data set which represent important features of that data set.
- ▶ Two very useful summary statistics which give numerical values for the centre of a distribution are the **mean** and the **median**.
- ▶ The **mean** is denoted \bar{x} and the mean = $\frac{\text{sum of data values}}{\text{total number of data values}}$

$$= \frac{\sum x}{n}$$
- ▶ The **median** is the middle value in the ordered data, which is the $\left(\frac{n+1}{2}\right)$ th observation from the end of the list when n is odd, or the average of the $\frac{n}{2}$ th and the $\left(\frac{n}{2} + 1\right)$ th observations when n is even.
- ▶ When the distribution is skewed, the **median** is the preferred measure of centre.

Exercise 2E

Building understanding

Example 16

- 1 For the following data set:

1 2 1 0 2 3 1 1 2 6 7

- a Find the sum of the data values.
- b Hence, find the mean of the data set.

Example 17

- 2 For the following data set:

21 12 35 53 32 63 11 19 62 17 24 34 95

- a Order the data from smallest to largest value.
- b Hence, find the median of the data set.

Developing understanding

- 3 Find, without using a calculator, the mean 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 18

- 4 Find, without using a calculator, the median for each of these ordered data sets.

- 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

- 5** Without a calculator, find the median of the data displayed in the following stem plots.

a *Monthly rainfall (mm)*

4|8 represents 48 mm

```

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

```

b *Battery time (hours)*

1|7 represents 17 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

```

- 6** The following data 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.



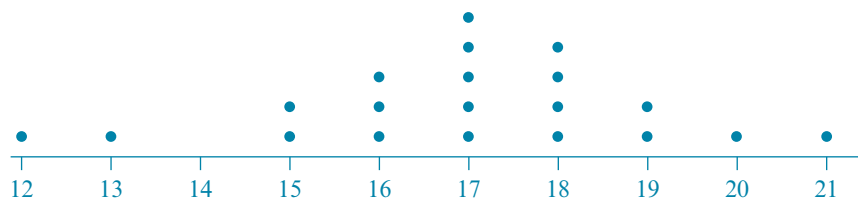
- 7** Find the mean of each of the data sets displayed in the stem plots in Question 5.
- 8** The prices, in dollars, of apartments sold in a particular suburb during one month were:

\$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 for this data set? Explain your answer.

- 9** Find the mean and median of the data set displayed in the following dot plot.



- 10** Suppose that the mean of a data set with 8 values is 23.6, and that when another data value is added, the mean becomes 24.9. What is the data value that has been added to the data set?

Testing understanding

11 The mean score for Mr Miller's class in the Mathematics test was 67. The mean score for Mrs Lacey's class was 72. If 23 students took the test in Mr Miller's class, and 20 took the test in Mrs Lacey's class, what was the mean score across both classes?

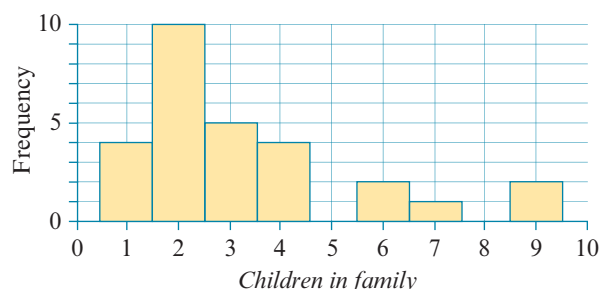
12 Jessie is completing a science experiment. She has 20 data values and has calculated the mean of the data set to be 15.6 and the median to be 15.0. When she is typing up her results, she mistypes the smallest data value, typing 1.6 instead of 10.6.

a What would the new mean of the data set be?

b What would the new median of the data set be?

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

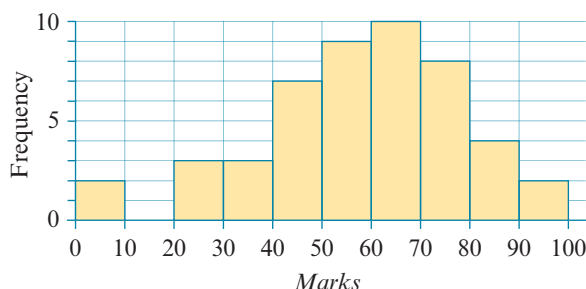
a What is the median number of children in the family for this class?



b What is the mean number of children in the family for this class?

14 The following histogram gives the scores on a general knowledge quiz for a class of Year 11 students.

a What can you say about the median mark on the quiz?



b Can you identify the mean mark from this histogram? Justify your answer.

15 The following percentage frequency table gives the heights of members of a sporting team:

a What can you say about the median height of the team members?

b Can you identify the mean height from this percentage frequency table? Justify your answer.

Height (cm)	Percentage Frequency
170.0–174.9	5
175.0–179.9	22
180.0–184.9	35
185.0–189.9	26
190.0–194.9	10
195.0–199.9	2
Total	100

2F Measures of spread

Learning intentions

- ▶ To be able to understand the range, interquartile range and standard deviation as measures of spread.
- ▶ To be able to know when to use each as a measure of spread for a particular distribution.
- ▶ To be able to learn to use a CAS calculator to calculate summary statistics.

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.

$$R = \text{largest data value} - \text{smallest data value}$$



Example 19 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.

Explanation

For Task A, the minimum mark is 2 and the maximum mark is 94.

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

Solution

Range for Task A = $94 - 2 = 92$

Range for Task B = $91 - 11 = 80$

Now try this 19 Finding the range (Example 19)

The weights of 19 cats are displayed in this ordered stem plot. Find the range.

Weight (kg)	key: 1 2 represents 1.2 kg
0	5 9
2	1 3 5 8
3	0 0 4 9 9
4	0 4 5 8
5	3 7
6	3 4

Hint 1 Make sure that you carefully look at the key to the data values.

In Example 19, 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 key: 1 | 2 represents 12 marks

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 key: 1 | 2 represents 12 marks

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

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: $IQR = Q_3 - Q_1$.

We can interpret the interquartile range as follows:

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

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.

Note that the *median* which has 50% of the data below and 50% of the data values above, is denoted as Q_2 , but we don't commonly use this notation.



Example 20 Finding the interquartile range (IQR)

Find the interquartile range for Task A and Task B in Example 19 and compare.

Explanation

- 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.
- 2 The median of the upper half (Q_3) is the 8th value.
- 3 Determine the IQR.
- 4 Repeat the process for Task B.
- 5 Compare the IQR for Task A to the IQR for Task B.

Solution

Task A

Lower half:

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

$Q_1 = 22$

Upper half:

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

$Q_3 = 47$

$IQR = Q_3 - Q_1 = 47 - 22 = 25$

Task B

$Q_1 = 31$

$Q_3 = 73$

$IQR = Q_3 - Q_1 = 73 - 31 = 42$

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

Now try this 20**Finding the interquartile range (IQR) (Example 20)**

Find the interquartile range of the weights of the 19 cats whose weights are displayed in this ordered stem plot.

Weight (kg)	1 2 represents 1.2 kg
0 5 9	
2 1 3 5 8	
3 0 0 4 9 9	
4 0 4 5 8	
5 3 7	
6 3 4	

Hint 1 Locate the median.

Hint 2 Since there are an uneven number of values, omit the median. The upper and lower halves of the data will have 9 data values in each.

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 for any distribution, whether skewed or symmetric.

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 theoretical reasons (not important here) we average by dividing by $n - 1$, not n .

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 21 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.

Explanation

- 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.
- 2** First find the mean (\bar{x}) and then complete the table as shown.
- 3** Substitute the required values into the formula and evaluate.

Solution

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

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

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

So, the standard deviation is 1.

Now try this 21 Calculating the standard deviation (Example 21)

Use the formula above to calculate the standard deviation for this data set:

0 1 3 5.

Hint 1 Set up a table as in Example 21, with 4 rows for the data values.

Hint 2 Remember when any negative number is squared, the result is positive.

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 find 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

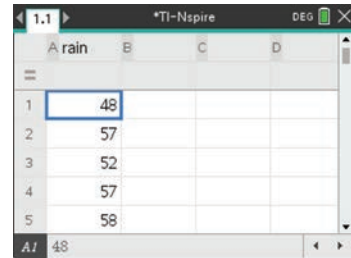
Determine the mean and standard deviation, median and interquartile range, and range.

Steps

- 1 Start a new document: Press $\boxed{\text{ctrl}} + \boxed{\text{on}}$ and select **New** (or press $\boxed{\text{ctrl}} + \boxed{\text{N}}$).

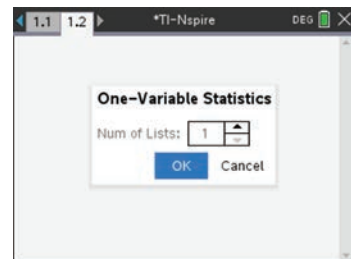
- 2 Select **Add Lists & Spreadsheet**.

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



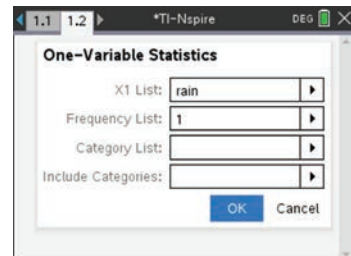
- 3 Press $\boxed{\text{ctrl}} + \boxed{\text{doc}}$ and select **Add Calculator** (or press $\boxed{\text{ctrl}} + \boxed{\text{on}}$ and arrow to $\boxed{+/-}$ and press $\boxed{\text{enter}}$).

- a Press $\boxed{\text{menu}} > \text{Statistics} > \text{Stat Calculations} > \text{One-Variable Statistics}$



- b Press $\boxed{\text{enter}}$

- c Use the $\boxed{\text{tab}}$ key to highlight **OK** and press $\boxed{\text{enter}}$ to generate the statistical results screen below.



OneVar rain,1: stat.results	
"Title"	"One-Variable Statistics"
" \bar{x} "	55.4167
" Σx "	665.
" Σx^2 "	37223.
" $s_x := s_{n-1}x$ "	5.80687
" $\sigma_x := \sigma_n x$ "	5.55965
"n"	12.
"MinX"	48.
" Q_1X "	49.5

" $s_x := s_{n-1}x$ "	5.80687
" $\sigma_x := \sigma_n x$ "	5.55965
"n"	12.
"MinX"	48.
" Q_1X "	49.5
"MedianX"	57.
" Q_3X "	59.
"MaxX"	67.
" $SSX := \Sigma(x-\bar{x})^2$ "	370.917

- 4 Write the answers rounded to one decimal place.

$$\bar{x} = 55.4, S = 5.8 \quad M = 57$$

$$\text{IQR} = Q_3 - Q_1 = 59 - 49.5 = 9.5$$

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

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

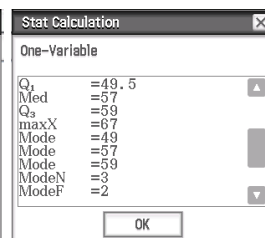
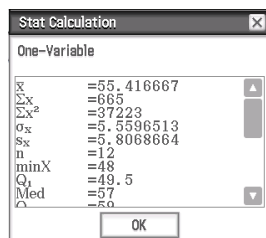
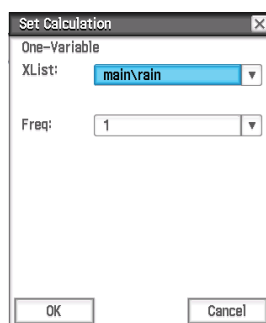
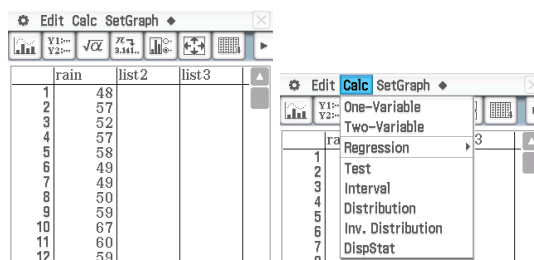
Determine the mean and standard deviation, median and interquartile range, and range.

Steps

- 1 Open the **Statistics** application and enter the data into the column labelled *rain*.
- 2 To calculate the mean, median, standard deviation and quartiles:
 - Select **Calc** from menu bar.
 - Tap **One-Variable**.
- 3 Complete the Set Calculation dialog box. For:
 - **XList:** select **main \ rain** (▼)
 - **Freq:** leave as **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 for additional statistics if required.
- 5 Write the answers rounded to one decimal place.



$$\bar{x} = 55.4, S = 5.8, M = 57$$

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

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

Section Summary

- ▶ The **range** of a data set can be used to indicate the spread of the data set. The range is the difference between the largest and smallest values in the data set.

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

- ▶ The **quartiles** of a data set are values which divide the data set into quarters. The **first quartile** is the median of the lower half of the observations, and the **third quartile** is the median of the upper half of the observations.
- ▶ The **interquartile range or IQR** is a measure of spread of a data set

$$\text{IQR} = Q_3 - Q_1$$

- ▶ The **standard deviation** is another measure of spread of a data set. It measures the variation of the data values around the mean.
- ▶ When the distribution is skewed or has outliers, the **interquartile range** is the preferred measure of spread.
- ▶ After entering data into a CAS spreadsheet, the mean, standard deviation, minimum, maximum, Q_1 and Q_3 can easily be determined.

Exercise 2F

Building understanding

Example 19

- 1 For the following data set:

1 2 1 0 2 3 1 2 6 7

- a Order the data from smallest to largest value.
- b Find the minimum and maximum values, and hence find the value of the range.
- c Find the value of Q_1 , the median of the lower half of the data values.
- d Find the value of Q_3 , the median of the upper half of the data values.
- e Hence, find the value of the interquartile range (IQR).

Example 21

- 2 For the following data set:

1 2 3 6

- a Find the value of the mean.
- b Use a table to calculate the value of the standard deviation, using the formula:

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

Developing understanding

- 3** Find, without using a calculator, the IQR and range of each of these ordered data sets.

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

- 4** Without a calculator, determine the IQR for the data displayed in the following stem plots.

a *Monthly rainfall (mm)*

key: 4|8 represents 48 mm

```

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

b *Battery time (hours)*

key: 1|7 represents 17 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
4 | 0 1 6
  
```

- 5** Without using a calculator, find the median and quartiles for the data displayed in the following dot plot:



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

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

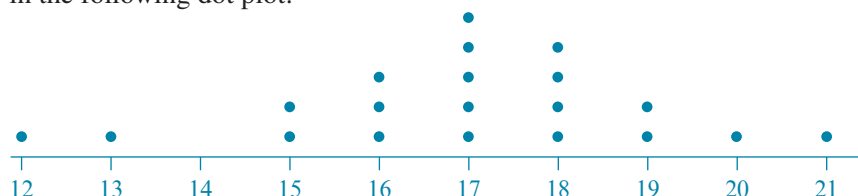
Find the mean and standard deviation, median and IQR, and range for the volume of drink in the cans. Give answers rounded to one decimal place.

- 7** The serum cholesterol levels for a sample of 18 people are:

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

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

- 8 Find the range, interquartile range and standard deviation of the data set displayed in the following dot plot:



- 9 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.

Testing understanding

- 10 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 6.1 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 6.1 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**.

2G Percentages of data lying within multiple standard deviations of the mean

Learning intentions

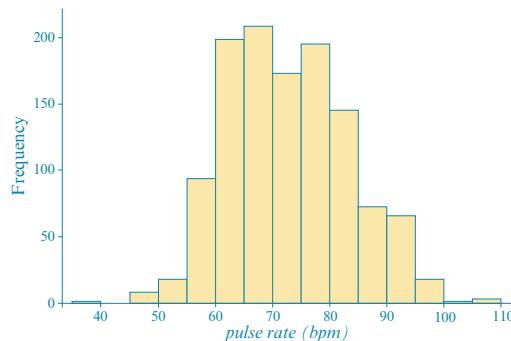
- To be able to consider the percentage of data lying within several standard deviations of the mean for a range of distributions.
- To be able to apply our knowledge of the mean and standard deviation to understand a data distribution.
- To be able to introduce the 68 - 95 - 99.7% rule for symmetric and bell shaped data distributions.

In the previous section we defined the interquartile range, IQR, and could readily interpret this statistic as the spread of the middle 50% of the data values in a distribution.

We also defined the values of the mean and the standard deviation for a distribution. Can we use these two statistics in combination to tell us a bit more about the distribution of the random variables we are exploring?

It turns out that when the data distribution is **symmetric** and approximately **bell shaped**, we can estimate the percentage of the data lying within several standard deviations of the mean.

We will explore this idea using data collected from 1000 people for the variable *pulse rate*, measured in beats per minute, displayed in the following histogram.



From the histogram we can see that this distribution is approximately symmetric and bell shaped.

From the data, the mean *pulse rate* is $\bar{x} = 72.31$ and the standard deviation is $s = 10.29$. We can use this information to construct intervals which are one, two and three standard deviations either side from the mean, as follows:

One SD: $(\bar{x} - s, \bar{x} + s) = (72.31 - 10.29, 72.31 + 10.29) = (62.02, 82.60)$

Two SD: $(\bar{x} - 2s, \bar{x} + 2s) = (72.31 - 2 \times 10.29, 72.31 + 2 \times 10.29) = (51.73, 92.89)$

Three SD: $(\bar{x} - 3s, \bar{x} + 3s) = (72.31 - 3 \times 10.29, 72.31 + 3 \times 10.29) = (41.11, 103.18)$

We can now go back to the original 1000 data values and determine the percentage of the data which lies within each of the three intervals. When we do this we find that for this particular set of 1000 values of *pulse rate*:

- 68% of the data values lie within 1 standard deviation of the mean
- 95% of the data values lie within 2 standard deviations of the mean
- 99.7% of the data values lie within 3 standard deviations of the mean.

If we repeat this analysis for another set of values for *pulse rate*, we find that we get very similar results. In fact, we are able to show theoretically that there is a general rule which can be applied here.

The 68 - 95 - 99.7% rule

For data distribution which is approximately symmetric and bell shaped, approximately:

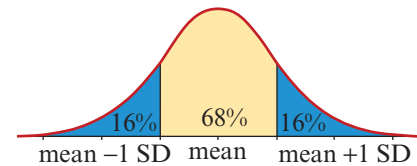
- 68% of the data will lie within one standard deviation of the mean.
- 95% of the data will lie within two standard deviations of the mean.
- 99.7% of the data will lie within three standard deviations of the mean.

Of course, to be able to show that these rules apply empirically we would require a large data set (preferably at least 200 values).

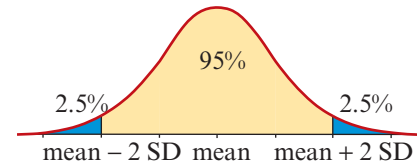
The 68–95–99.7% rule in graphical form

It is helpful to also illustrate this property of the standard deviation graphically. In the following diagrams a smooth curve has been used to imply that the underlying data distribution is generally symmetric and bell shaped.

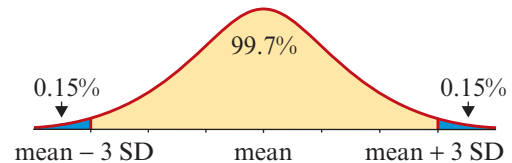
- Around 68% of the data values will lie within one standard deviation (SD) of the mean.



- Around 95% of the data values will lie within two standard deviations of the mean.



- Around 99.7% of the data values will lie within three standard deviations of the mean.




Example 22 Percentages of data lying within 1, 2 or 3 standard deviations of the mean

The distribution of the examination scores for a very large statewide examination is approximately symmetric and bell shaped, with a mean of 65 and with a standard deviation of 10.

- a** Approximately what percentage of students scored between 55 and 75?
- b** Approximately what percentage of students scored between 45 and 85?
- c** Approximately what percentage of students scored between 35 and 95?

Explanation

- a** A score of 55 is 1SD below the mean of 65 and a score of 75 is 1SD above the mean.
- b** A score of 45 is 2SD below the mean of 65 and a score of 85 is 2SD above the mean.
- c** A score of 35 is 3SD below the mean of 65 and a score of 95 is 3SD above the mean.

Solution

Approximately 68% of the scores are between 55 and 75.

Approximately 95% of the scores are between 45 and 85.

Approximately 99.7% of the scores are between 35 and 95.


Example 23 Finding the interval for a given percentage

The distribution of the diameter of bolts produced in a factory is approximately symmetric and bell shaped, with a mean of 5 mm and with a standard deviation of 0.01 mm.

- a** If approximately 68% of the bolts measure between a and b , what are possible values for a and b ?
- b** If approximately 95% of the bolts measure between c and d , what are possible values for c and d ?
- c** If approximately 99.7% of the bolts measure between e and f , what are possible values for e and f ?

Explanation

- a** The interval which contains 68% of the bolts is 1SD either side of the mean.
- b** The interval which contains 95% of the bolts is 2SD either side of the mean.
- c** The interval which contains 99.7% of the bolts is 3SD either side of the mean.

Solution

$$a = 5 - 0.01 = 4.99 \text{ mm}$$

$$b = 5 + 0.01 = 5.01 \text{ mm}$$

$$c = 5 - 2 \times 0.01 = 4.98 \text{ mm}$$

$$d = 5 + 2 \times 0.01 = 5.02 \text{ mm}$$

$$e = 5 - 3 \times 0.01 = 4.97 \text{ mm}$$

$$f = 5 + 3 \times 0.01 = 5.03 \text{ mm}$$

Section Summary

- ▶ Knowing the mean and standard deviation of a distribution allows us to make predictions about the percentage of the data that lies within specific intervals.
- ▶ If the data distribution is symmetric and bell shaped then approximately:
 - ▶ 68% of the data will lie within one standard deviation of the mean.
 - ▶ 95% of the data will lie within two standard deviations of the mean.
 - ▶ 99.7% of the data will lie within three standard deviations of the mean.

**Exercise 2G****Building understanding**

- 1 Suppose that the mean of a large data set is 15.8, and the standard deviation is 2.3. Find the interval which is:
 - a One standard deviation either side of the mean.
 - b Two standard deviations either side of the mean.
 - c Three standard deviations either side of the mean.
- 2 Suppose that the mean of a large data set is 435.6, and the standard deviation is 53.3. Find the interval which is:
 - a One standard deviation either side of the mean.
 - b Two standard deviations either side of the mean.
 - c Three standard deviations either side of the mean.

Developing understanding**Example 22**

- 3 Suppose the distribution of height for females in a certain country is approximately symmetric and bell shaped, with a mean of 163 cm and a standard deviation of 8 cm.
 - a Approximately what percentage of females are between 147 cm and 179 cm tall?
 - b Approximately what percentage of females are between 139 cm and 187 cm tall?
- 4 The distribution of IQ scores in a certain country is approximately symmetric and bell shaped, with a mean of 100 and a standard deviation of 15.
 - a Approximately what percentage of people have an IQ score between 55 and 145?
 - b Approximately what percentage of people have an IQ score between 70 and 130?
- 5 The distribution of weights of eggs from a farm is approximately symmetric and bell shaped, with a mean of 60 gm and a standard deviation of 3 gm.
 - a Approximately what percentage of eggs have a weight between 57 gm and 63 gm?
 - b Approximately what percentage of eggs have a weight between 51 gm and 69 gm?

- 6** The distribution of times taken for people to solve a puzzle is approximately symmetric and bell shaped, with a mean of 24 seconds and a standard deviation of 4 seconds.
- a** Approximately what percentage of people take between 20 seconds and 28 seconds to solve the puzzle?
 - b** Approximately what percentage of people take between 16 seconds and 32 seconds to solve the puzzle?

Example 23

- 7** The distribution of the volume of soft drink in a 1 litre bottle is approximately symmetric and bell shaped, with a mean of 1.0 litre and a standard deviation of 2 mL. If approximately 95% of the bottle contains between a litres and b litres, what are possible values for a and b ?
- 8** The distribution of salaries in a certain country is approximately symmetric and bell shaped, with a mean of \$91 000 per year and a standard deviation of \$26 500.
- a** If approximately 95% of people have salaries between a and b , what are possible values for a and b ?
 - b** If approximately 99.7% of people have salaries between c and d , what are possible values for c and d ?

Testing understanding

- 9** Suppose the number of hours of exercise per week undertaken by people in a certain country is approximately symmetric and bell shaped, with a mean of 6.2 hours and a standard deviation of 1.6 hours.
- a** If approximately 95% of people exercise between a and b hours per week, what are possible values for a and b ?
 - b** If 99.7% of people exercise between c and d hours per week, what are possible values for c and d ?
 - c** If approximately 50% of people exercise for more than e hours per week, what is the value of e ?
 - d** Approximately what percentage of people exercise for between 7.8 and 9.4 hours per week?

2H Boxplots

Learning intentions

- ▶ To be able to introduce the boxplot as a plot for displaying the distribution of a numerical data.
- ▶ To be able to introduce an exact definition of an **outlier**.
- ▶ To be able to define a **simple boxplot** and a **boxplot with outliers**.
- ▶ To be able to determine the characteristics of centre and spread from a boxplot.

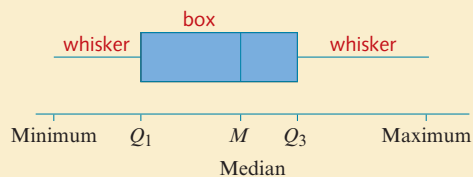
We already have three plots which we can use to display the distribution of a numerical variable, namely the histogram, the dot plot, and the stem and leaf plot. In this section we introduce another plot for displaying numerical data, the **boxplot**. The boxplot is extremely useful as it allows us to summarise quite large data sets into concise plots.

The simple boxplot

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 set, 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 24 Constructing a boxplot from a five-number summary

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

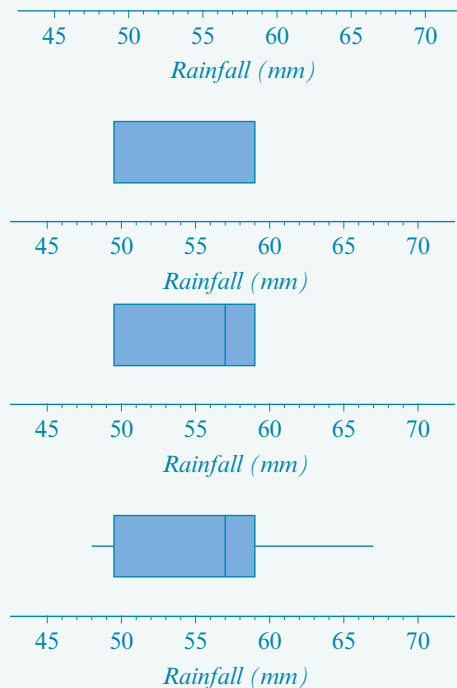
Month	J	F	M	A	M	J	J	A	S	O	N	D
Rainfall (mm)	48	57	52	57	58	49	49	50	59	67	60	59

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

$$\text{Min} = 48 \quad Q_1 = 49.5 \quad M = 57 \quad Q_3 = 59 \quad \text{Max} = 67$$

Explanation

- 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 $Q_3 = 59$.
- 3 Mark in the median value with a vertical line segment at $M = 57$.
- 4 Draw in the whiskers, which are lines joining the midpoint of the ends of the box to the minimum and maximum values: 48 and 67, respectively.

Solution

Now try this 24 Constructing a boxplot from a five-number summary (Example 24)

Construct a boxplot to display the following five-number summary:

$$\text{Min} = 7 \quad Q_1 = 14 \quad M = 19 \quad Q_3 = 24 \quad \text{Max} = 34$$

Hint 1 Make sure the scale on the number line spans the maximum and minimum values.

Hint 2 Locate each value of the five-number summary with a dot before you construct the boxplot.

Boxplots with outliers

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

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

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

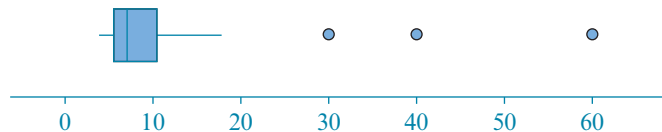
By modifying the boxplots, we can decide which explanation is most likely, but first we need a more exact definition of an outlier.

Defining outliers

Outlier

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

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



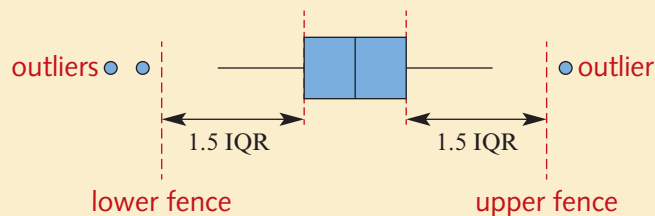
Upper and lower fences

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 times the interquartile range (or box widths) above and below the ends of the box. Data values outside these fences are classified as possible outliers and plotted separately. Note that if a data point lies exactly on an upper or lower fence, then it is not considered an outlier.

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 25 Constructing a boxplot showing outliers

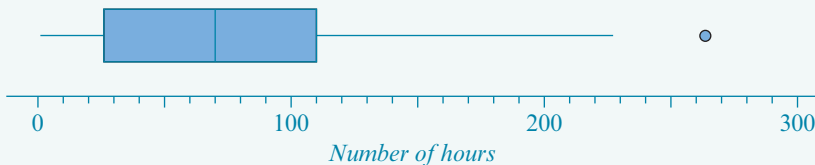
The number of hours that each of 33 students spent on a school project is shown below.

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.

Explanation

- 1 From the ordered list, state the minimum and maximum values. Find the median, the $\frac{1}{2}(33 + 1)$ th = 17th value.
- 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.
- 3 Determine the IQR.
- 4 Determine the upper and lower fences.
- 5 Locate any values outside the fences, and the values that lie just inside the limits (the whiskers will extend to these values).
- 6 The boxplot can now be constructed as shown below. The upper whisker extends to the second highest value, and the circle denotes the outlier. The fences are not shown on the boxplot.



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

Solution

Minimum = 2 hours

Maximum = 264 hours

Median = 71 hours

$$\text{First quartile, } Q_1 = \frac{24 + 27}{2} = 25.5$$

$$\text{Third quartile, } Q_3 = \frac{108 + 111}{2} = 109.5$$

$$IQR = Q_3 - Q_1 = 109.5 - 25.5 = 84$$

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

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

There is one outlier: 264 hours.

The largest value that is not an outlier is 226 hours.

Now try this 25 Constructing a boxplot showing outliers (Example 25)

The number of hours that each of a sample of 22 people spent in paid employment last week is shown here.

32 16 40 20 35 40 40 43 40 40 35
45 40 72 75 30 60 60 40 55 48 40

Construct a boxplot with outliers for this data set.

Hint 1 Start by putting the data in order, and then determining the five-number summary.

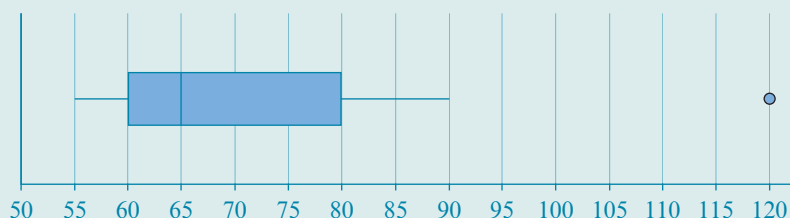
Hint 2 Determine the IQR, and hence the values for the lower and upper fences.

Hint 3 Remember the whiskers join the ends of the box to the smallest and largest values which are not outliers. They do not extend to the upper and lower fences.

We can use a boxplot to tell us a lot about the distribution of a data set, as is shown in the following example.

**Example 26** Estimating percentages from a boxplot

For the boxplot shown, estimate the percentage of values which are:



- a** less than 60 **b** less than 65 **c** more than 80
d between 60 and 80 **e** between 60 and 120

Explanation

a 60 is the first quartile (Q_1).

b 65 is the median (M , or sometimes Q_2).

c 80 is the third quartile (Q_3).

d 75% of the data values are less than 80, and 25% of the data values are less than 60.

e 100% of the data values are less than 120, and 25% of the data values are less than 60.

Solution

25% of the data values are less than 60.

50% of the data values are less than 65.

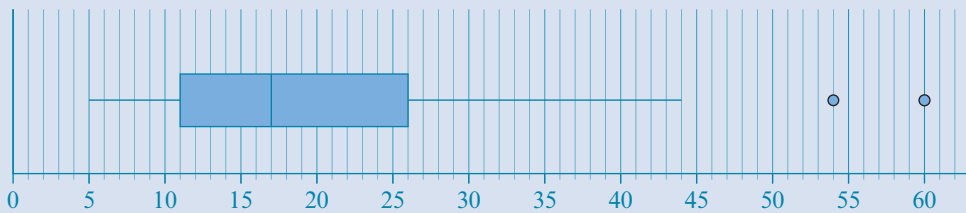
75% of the data values are less than 80, so 25% of the data values are greater than 80.

Thus 50% of the values are between 60 and 80.

Thus 75% of the values are between 60 and 120.

Now try this 26**Estimating percentages from a boxplot (Example 26)**

For the boxplot shown, estimate the percentage of values which are:



- a** less than 26 **b** less than 5 **c** less than 11
d between 11 and 26 **e** between 5 and 26

Hint 1 Start by identifying the values of the five-number summary from the boxplot.

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.

How to construct a boxplot using the TI-Nspire CAS

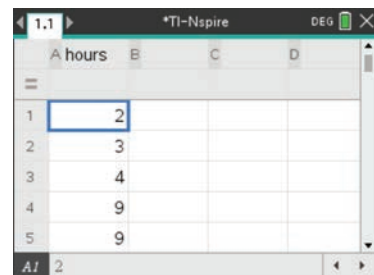
The number of hours that each of 33 students spent on a school project is shown below.

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.

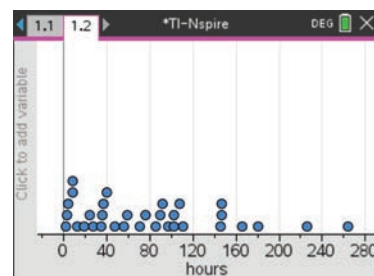
Steps

- Press and select **New** (or use **ctrl** + **N**).
- Select **Add Lists & Spreadsheet**.
Enter the data into a list called *hours* as shown.



- Statistical graphing is done through the **Data & Statistics** application. Press **ctrl** + **doc** and select **Add Data & Statistics** (or press , 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.



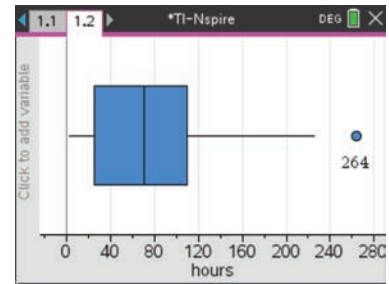
- a Press **[tab]** to show the list of variables. Select the variable *hours*. Press **[enter]** to paste the variable *hours* to that axis. A dot plot is displayed as the default plot.
- b To change the plot to a boxplot press **[menu]>Plot Type>Box Plot**, then **[enter]** or 'click' (press **[2nd][<] [2nd][>]**). Outliers are indicated by a dot(s).

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: **Q₁ = 25.5**
- median is 71: **Median = 71**
- third quartile is 109.5: **Q₃ = 109.5**
- maximum value is 264: **maxX = 264**. It is also an outlier.



How to construct a boxplot using the ClassPad

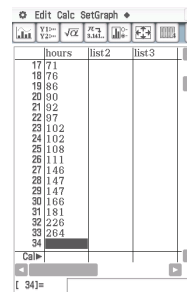
The number of hours that each of 33 students spent on a school project is shown below.


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 boxplot 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*.





- 2 Open the **Set StatGraphs** dialog box by tapping  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  to exit.



- 3 Tap  to plot the boxplot.
- 4 Tap  to obtain a full-screen display.

Note: In the screen shot shown, the window parameters were adjusted to display vertical grid lines, no scale along the y-axis and less space below the x-axis. This was achieved by tapping on  and selecting 20 for the x scale, 0 for the y scale and reducing the ymin value.

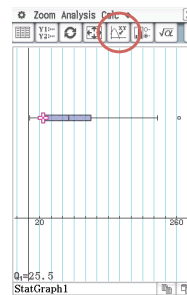
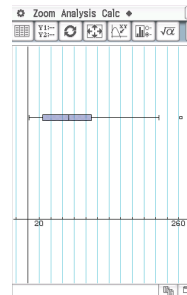
- 5 Key values can be read from the boxplot by tapping .

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.



Section Summary

- ▶ A boxplot is a useful plot for displaying the distribution of a numerical data set.
- ▶ A **simple boxplot** is a pictorial representation of the five-number summary.
- ▶ The five-number summary consists of the minimum, the first quartile Q_1 , the median, the third quartile Q_3 and the maximum.
- ▶ From the boxplot we can see:
 - ▶ 25% of the data is between the minimum and Q_1 .
 - ▶ 25% of the data is between Q_1 and the median.
 - ▶ 25% of the data is between the median and Q_3 .
 - ▶ 25% of the data is between Q_3 and the maximum.
- ▶ A **boxplot with outliers** shows the five-number summary as well as any **outliers**.
- ▶ The **lower fence** is $Q_1 - 1.5 \times \text{IQR}$.
- ▶ The **upper fence** is $Q_3 + 1.5 \times \text{IQR}$.
- ▶ Any data value less than the lower fence, or greater than the upper fence, is shown on the boxplot as an outlier.



Exercise 2H

Building understanding

Example 24

- 1** The five-number summary for a data set is:

$$\text{Min} = 5 \quad Q_1 = 10 \quad M = 20 \quad Q_3 = 25 \quad \text{Max} = 45$$

- a** Construct a number line starting at 0, ending at 50, and marked off in units of 10.
- b** Mark in each of the values of the five-number summary on the number line.
- c** Hence construct a simple boxplot.

Developing understanding

- 2** The data shows how many hours each of a group of forty-one players spent at training 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

- a** Find the five-number summary for this data.
b Use this five-number summary to construct a simple boxplot by hand.

- 3** The five-number summary for a data set is:

Min = 0 $Q_1 = 13$ $M = 16$ $Q_3 = 19$ Max = 35

Determine the values of the lower and upper fences.

Example 25

- 4** The five-number summary for a data set is:

Min = 14 $Q_1 = 45$ $M = 55$ $Q_3 = 65$ Max = 99

- a** Determine the values of the upper and lower fences.
b The smallest three values in the data set are 14, 18, 34 and the largest are 90, 94, 99. Which of these are outliers?

- 5** 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

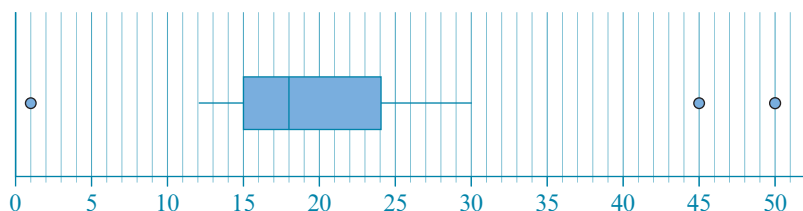
The five-number summary is:

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

- a** Use the information from the five-number summary to determine the values of the lower and upper fences.
b Without using a calculator, determine the value(s) of any outliers.
c Without using a calculator, construct a boxplot showing any outliers.

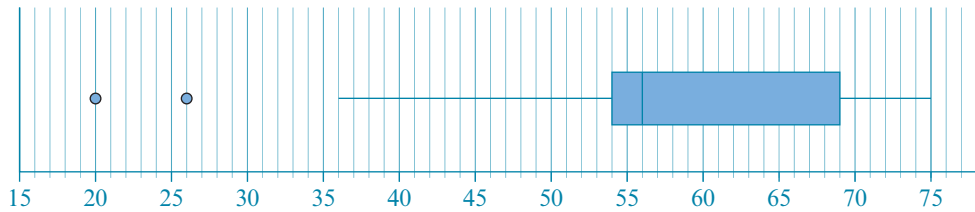
Example 26

- 6** Use the boxplot below to estimate the percentage of values that are:



- a** less than 18 **b** less than 12 **c** between 12 and 24 **d** more than 24

- 7** The boxplot below displays the scores on an exam for a group of students. Use it to estimate the percentage of students who had scores that are:



- a** less than 69 **b** less than 54 **c** between 54 and 69
d more than 75 **e** between 20 and 69 **f** between 56 and 69

- 8** 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 a boxplot.

- 9** The times, in seconds, that 35 children took to tie 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	

- a** Use a CAS calculator to construct a boxplot.
b If 25% of the children took less than k seconds to tie their shoelaces, what is the value of k ?

- 10** 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.
c How many books were borrowed by the top 25% of library users?

- 11** The following table gives the prices for houses 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

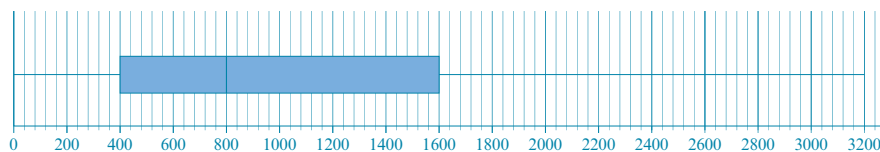
- Use a CAS calculator to construct a boxplot of the data.
 - Use the boxplot to identify any possible outliers, and write down their values.
 - What is the price range of the top 50% most expensive houses?
- 12** The time taken, in seconds, for a group of children to complete a puzzle is:

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	

- Use a CAS calculator to construct a boxplot of the data.
- Use the boxplot to identify any possible outliers, and write down their values.
- What is the slowest time for the fastest 25% of children?

Testing understanding

- 13** The following boxplot summarises the weekly income (in dollars) for a large sample of people. Use the boxplot to answer the following questions.



- What is the approximate value of the median income?
- What is the approximate value of the interquartile range of incomes?
- What is the minimum salary for a person who is in the top 25% of earners?
- The maximum weekly income for this sample of people is \$3200. Confirm that it is not an outlier.

- 14** The table shows the percentage of people using the internet in 23 countries in 2020.

Country	Internet users (%)	Country	Internet users (%)
Afghanistan	18.8	Malaysia	67.5
Argentina	78.6	Morocco	61.6
Australia	85.1	New Zealand	86.6
Brazil	70.2	Saudi Arabia	88.6
Bulgaria	53.1	Singapore	82.0
China	59.3	Slovenia	72.7
Colombia	63.2	South Africa	53.1
Greece	59.9	United Kingdom	94.7
Hong Kong (China)	80.5	United States	88.5
Iceland	96.5	Venezuela	61.5
India	40.6	Vietnam	66.3
Italy	92.9		

- Use a CAS calculator to construct a boxplot of the data.
- Use the boxplot to identify any possible outliers, and write down their values.
- What is the minimum internet usage for the top 75% of countries?

2I Comparing the distribution of a numerical variable across groups

Learning intentions

- ▶ To be able to use back-to-back stem plots to compare the distributions of two numerical variables.
- ▶ To be able to use parallel boxplots to compare the distributions of two or more numerical variables.
- ▶ To be able to write a report which communicates these comparisons.

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 gymnastics 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 **parallel boxplots**.

Comparing distributions using back-to-back stem plots

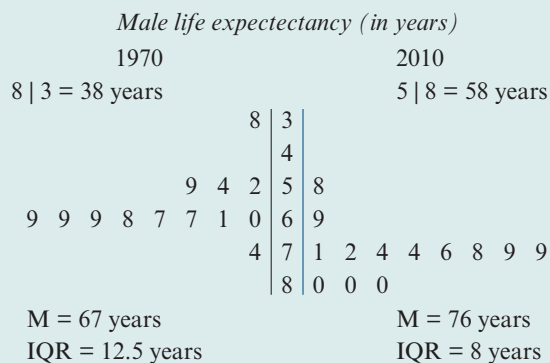
A back-to-back stem plot differs from the stem plots you have met in the past in that it has a single stem with two sets of leaves, one for each of the two groups being compared.



Example 27 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, *Male life expectancy* is the numerical variable. *Year*, which takes the values 1970 and 2010, is the categorical variable.



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.

Explanation

- 1 Centre: Write a sentence using the medians to compare centres.

Solution

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

2 Spread: Write a sentence using the IQRs to compare spreads.

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

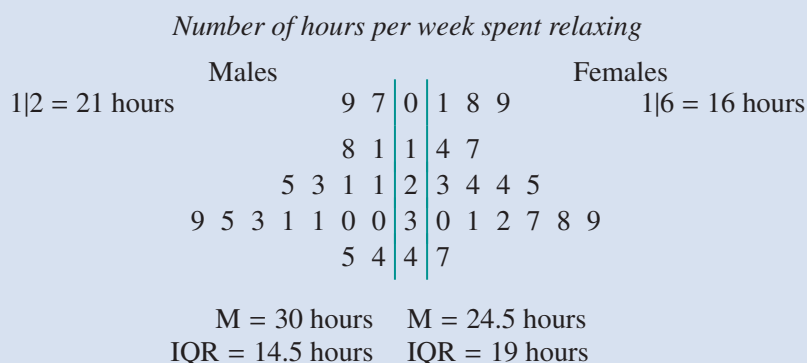
The spread of life expectancies of males in 2010 (IQR = 8 years) was lower than the spread in 1970 (IQR = 12.5 years).

In conclusion, the median life expectancy for men in these countries has increased over the last 40 years, and the variability in male life expectancy has decreased over this time interval.

Now try this 27

Comparing distributions using back-to-back stem plots (Example 27)

The following back-to-back stem plot displays the distributions of the number of hours per week spent relaxing by a group of 17 males and 16 females.



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.

Hint 1 Make sure that you **compare** the summary statistics using terms such as ‘more than’ or ‘less than’, don’t just state their values.

Hint 2 Ensure that your final comparison statement is clear and informative.

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, parallel boxplots can also be used when there are more than two groups.

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

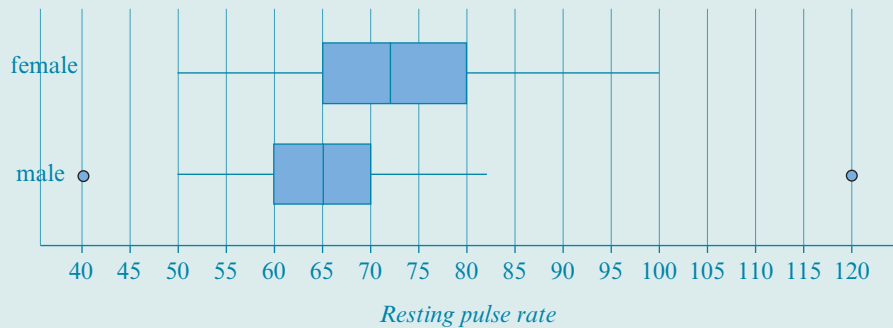
When comparing distributions of a numerical variable across two or more groups using parallel boxplots, the report should address the key features of:

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


Example 28 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.


Explanation

- 1** Centre: Determine values of the medians from the plot (the vertical lines in the boxes), and write a sentence comparing these values.
- 2** Spread: Determine the spread of the two distributions using IQRs (the widths of the boxes), and write a sentence comparing these values.
- 3** Outliers: Locate any outliers and write a sentence describing these.
- 4** Conclusion: Add a general conclusion based on these comparisons.

Solution

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

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

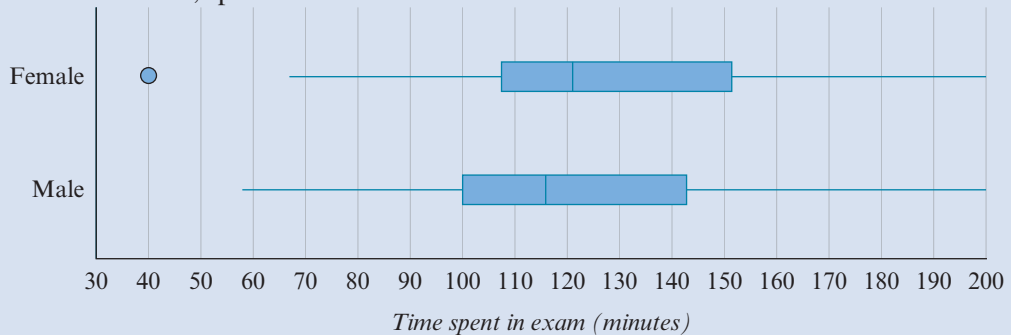
There are no female pulse rate outliers. The males with pulse rates of 40 and 120 were outliers.

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.

Now try this 28**Comparing distributions across two groups using parallel boxplots (Example 28)**

The following parallel boxplots display the distribution of time (in minutes) spent in an exam 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.



Hint 1 Make sure that you **compare** the summary statistics using terms such as 'more than' or 'less than,' don't just state their values.

Hint 2 Ensure that your final comparison statement is clear and informative.

**Section Summary**

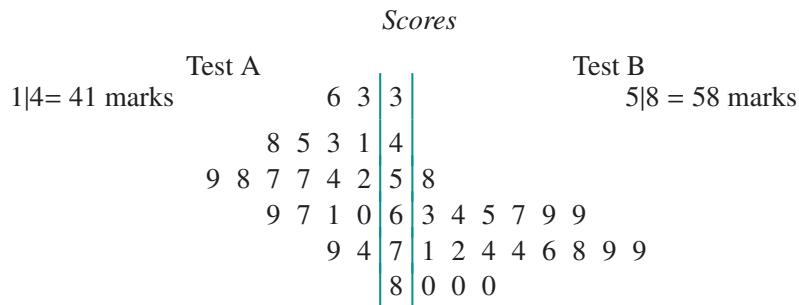
- ▶ Comparing numerical distributions across groups can be considered as investigating the association between a **numerical** variable and a **categorical** variable.
- ▶ Where there are only two groups, and the data sets are small, then **back-to-back stem plots** can be used to display the data.
- ▶ Where there are more than two groups, or the data sets are larger, then **parallel boxplots** can be used to display the data.
- ▶ The data distributions should be compared in terms of both **centre** and **spread**, quoting the median and IQR for each group.
- ▶ The values for any **outliers** should be mentioned.

Exercise 2I

Building understanding

Example 27

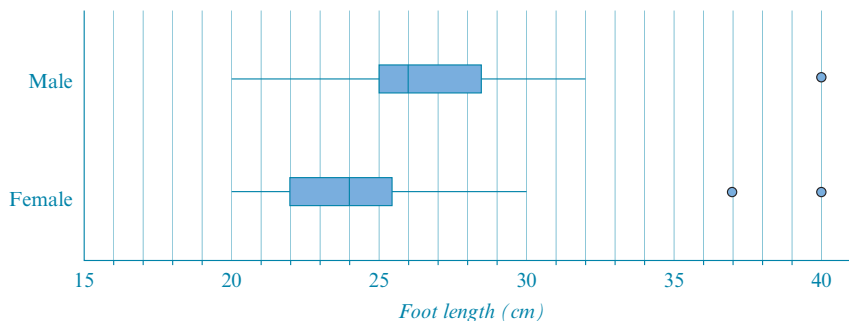
- 1 The following back-to-back stem plot displays the distribution of class scores for each of two tests (A and B).



- Find the median score for each test.
- Complete the sentence by choosing the correct alternative: 'The median score on Test A was (higher/lower) than the median score on Test B'.
- Find the IQR for each test.
- Complete the sentence by choosing the correct alternative: 'The scores on Test A were (more/less) variable than the scores on Test B'.

Example 28

- 2 The length (in cm) of the right foot for a group of 20 male and 20 female students in Year 9 are summarised in the following parallel boxplots.



- Complete the sentence by entering the values of the medians and choosing the correct alternative: 'The foot length for males (median = cm) is (longer/shorter) than the foot length for females (median = cm)'.
- Complete the sentence by entering the values of the IQR and choosing the correct alternative: 'The foot length for males (IQR = cm) is (more variable/less variable/similar in variability) to the foot length for females (IQR = cm)'.

Developing understanding

Comparing groups using back-to-back stem plots

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

Age females		Age males	
7 2 = 27 years	9	0	4 0 = 40 years
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		

- a** Calculate the median and the IQR for the ages of the females and males in this sample.
- b** Write a report comparing these distributions in terms of centre and spread.
- 4** 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.
- | Class B | | Marks | | Class A | |
|---------------------|-----|-------|---------------|------------------|--|
| 9 6 = 69 marks | 3 2 | 1 | 9 | 7 1 = 71 marks | |
| | | 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 | | |

- 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.
- 5** The following table shows the number of nights spent away from home in the past year by a group of 20 Australian tourists and by a group of 20 Japanese tourists:

Australian

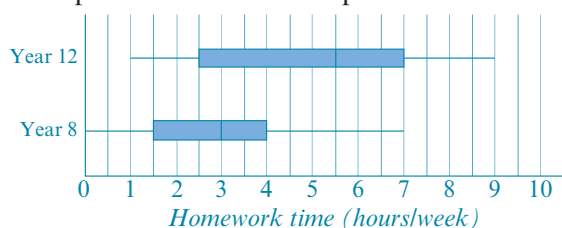
3	14	15	3	6	17	2	7	4	8
23	5	7	21	9	11	11	33	4	5

Japanese

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

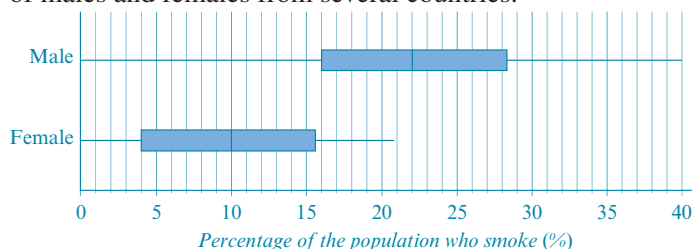
- a** Construct a back-to-back stem plot of these data sets.
- b** Determine the median and IQR for the two distributions.
- c** Write a report comparing the distributions of the number of nights spent away by Australian and Japanese tourists in terms of centre and spread.

- 6** The boxplots below display the distributions of homework time (in hours per week) of a sample of Year 8 and a sample of Year 12 students.



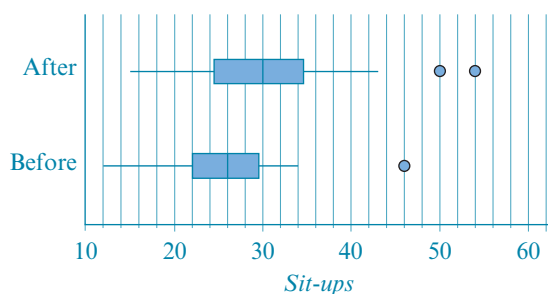
- a** Estimate the median and IQRs from the boxplots.
- b** Use the medians and IQRs to write a report comparing these distributions in terms of centre and spread in the context of the data.

- 7** The boxplots below display the distribution of smoking rates (in percentages) 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.

- 8** 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.

- 9** 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.

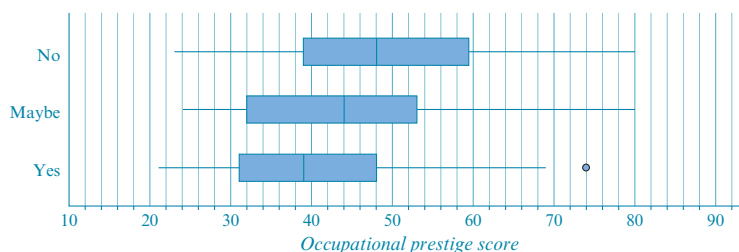
0% blood alcohol									
38	36	35	35	43	46	42	47	40	48
35	34	40	44	30	25	39	31	29	44

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.

Testing understanding

- 10** Occupational prestige is a numerical measure used to describe the respect that a particular occupation holds in a society. It is measured on a scale from 0 - 100. Researchers asked a sample of 332 people if they were thinking about changing to a different kind of work, to which they could respond yes, maybe or no. They also collected the occupational prestige scores for their current occupations. The data they collected is summarised in the following parallel boxplots:



- a** Use the information in the boxplots to write a report comparing the occupational prestige scores for each group with whether someone is thinking of changing jobs (yes, maybe or no) in terms of centre (medians), spread (IQRs) and outliers.
- b** Does there seem to be an association between occupational prestige score and whether someone is thinking of changing to a different kind of work?

Key ideas and chapter summary



Types of data	Data can be classified as categorical or numerical .
Categorical data	Categorical data arises when classifying or naming some quality or attribute. Categorical data can be nominal and ordinal .
Nominal data	Nominal data is a type of categorical data where the values of the variable are the names of groups.
Ordinal data	Ordinal data is a type of categorical data where there is an inherent order in the categories.
Numerical data	Numerical data arises from measuring or counting some quantity. Numerical data can be discrete or continuous .
Discrete data	Discrete data can only take particular numerical values, usually whole numbers, and often arises from counting.
Continuous data	Continuous data describes numerical data that can take any value, sometimes in an interval, and often arises from measuring.
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 as a percentage ; the percentage of times a value occurs.
Bar chart	A bar chart uses bars to display the frequency distribution of a categorical variable.
Mode, modal category/modal interval	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.
Histogram	A histogram uses columns 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, formed from the actual data values: suitable for small to medium-sized data sets.
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 middle of the distribution) and spread .

Summary statistics

Summary statistics are numerical values for special features of a data distribution such as centre and spread.

Mean

The **mean** (\bar{x}) is a summary statistic that can be used to locate the centre of a symmetric distribution. The value of the mean is determined from the formula: $\bar{x} = \frac{\sum x}{n}$

Range

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

Standard deviation

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

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

The 68-95-99.7% rule

For a data distribution which is approximately symmetric and bell shaped, approximately:

- 68% of the data will lie within one standard deviation of the mean.
- 95% of the data will lie within two standard deviations of the mean.
- 99.7% of the data will lie within three standard deviations of the mean.

Median

The **median** (M) is a summary statistic that can be used to locate the centre of a distribution. It is the midpoint of a distribution, so that 50% of the data values are less than this value and 50% are more. It is sometimes denoted as Q_2 .

Quartiles

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

Interquartile range

The **interquartile range (IQR)** gives the spread of the middle 50% of data values in an ordered data set. It is found by evaluating $IQR = Q_3 - Q_1$.

Five-number summary

The median, the first quartile and the third quartile, along with the minimum and the maximum values in a data set, are known as a **five-number summary**.

Outliers

Outliers are data values that appear to stand out from the rest of the data set. They are values that are less than the **lower fence** or more than the **upper fence**.

Lower and upper fences


The **lower fence** is equal to $Q_1 - 1.5 \times IQR$.
The **upper fence** is equal to $Q_3 + 1.5 \times IQR$.

Boxplot

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

Skills checklist



Download this checklist from the Interactive Textbook, then print it and fill it out to check your skills. 

2A 1 I can differentiate between nominal, ordinal, discrete and continuous data. ☐

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

- a The time between people arriving at the coffee shop.
- b The number of people in the queue at the coffee shop.
- c Customer's coffee preference (latte, cappuccino, black).
- d Customer's rating of the coffee (excellent, quite good, not that good).

2A 2 I can construct a frequency table and a percentage frequency table. ☐

e.g. Twenty students rated their school canteen as bad, ok or good, giving the following data: bad, bad, ok, ok, good, bad, good, ok, ok, ok, bad, ok, bad, good, good, good, good, bad, ok, ok.

Construct a percentage frequency table.

2A 3 I can identify the mode from a frequency table and interpret it. ☐

e.g. From the frequency table for the canteen rating (above), identify the mode.

2A 4 I can construct a bar chart from a frequency table. ☐

e.g. Construct a bar chart from the frequency table for the canteen rating (above).

2B 5 I can interpret and describe a frequency table and bar chart. ☐

e.g. Use the information in the table above to report on the students' ratings of their canteen.

2C 6 I can construct a histogram from raw data using a CAS calculator. ☐

e.g. The following data gives the number of games played in total by each member of an AFL club:

266	259	238	227	210	160	160	159	155	145	133
99	91	80	80	75	73	58	42	36	32	25
32	25	25	21	17	13	10	9	9	4	2

Use a CAS calculator to construct a histogram starting at 0 with column width 20.

2D 7 I can recognise symmetric, positively skewed and negatively skewed distributions. ☐

e.g. Describe the shape of the histogram of the number of games played (above).

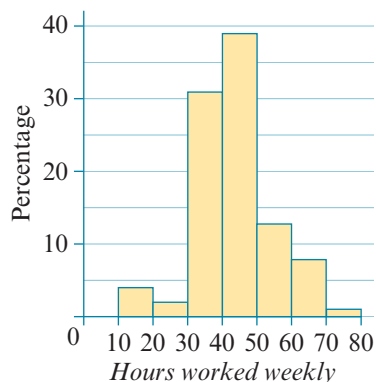
- 2D** **8** I can construct a dot plot from raw data. ☐
 e.g. Construct a dot plot of the following data:
 1 2 3 5 5 7 10 0 6 6 8 1 0 2 1 0 3 2 4 5
- 2D** **9** I can construct a stem plot from raw data. ☐
 e.g. Construct a stem plot of the ages of a group of people at a party:
 19 21 22 56 34 19 27 28 34 17
 66 37 20 20 21 45 26 19 23 29
- 2E** **10** I can determine the median and mean as measures of centre for a data set, and identify when it is more appropriate to use the median. ☐
 e.g. Determine the mean and median of the ages of the people at the party (from above). Which is the preferred measure of centre?
- 2F** **11** I can determine the quartiles of a data set and hence calculate the IQR. ☐
 e.g. Determine the quartiles and hence the IQR of the ages of the people at the party (from above).
- 2F** **12** I can find the standard deviation of a data set. ☐
 e.g. Find the standard deviation of the ages of the people at the party (from above).
- 2G** **13** I can use the 68-95-99.7% rule to estimate the percentages of a data distribution that are within one, two or three standard deviations of the mean. ☐
 e.g. Suppose the distribution of weights for a type of cat is approximately symmetric and bell shaped with a mean of 3.5 kg and a standard deviation of 0.5 kg. Approximately what percentage of cats weigh between 2 kg and 5 kg?
- 2H** **14** I can produce a five-number summary from a set of data and use it to construct a boxplot. ☐
 e.g. Complete the five-number summary for the ages of the people who attended the party (from above) and use it to construct a boxplot (without outliers).
- 2H** **15** I can determine the values of any outliers in a data set. ☐
 e.g. Calculate the values of the upper and lower fences, and use these to find any outliers in the ages of the people who attended the party (from above).
- 2I** **16** I can use a back-to-back stem plot or a boxplot to compare distributions in terms of centre, spread and outliers, and communicate this information in a written report. ☐
 e.g. See Example 27 and Example 28.

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 type of data generated is:
 - A variable
 - B numerical
 - C nominal
 - D ordinal
 - E discrete
- 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
- 3 For which of the following variables is a histogram an appropriate display?
 - A Hair colour
 - B Sex (male, female)
 - C Distance between towns (km)
 - D Postcode
 - E Weight (low, average, high)
- 4 In an experiment, researchers were interested in the effect of sunlight on the growth of plants. They exposed groups of plants to three levels of sunlight each day (3 - 5 hours, 6 - 8 hours, 9 - 11 hours) and then measured their growth in centimetres after three months. The variables *levels of sunlight* and *growth* are:
 - A both ordinal variables
 - B a numerical variable and an ordinal variable respectively
 - C an ordinal variable and a numerical variable respectively
 - D an ordinal variable and a nominal variable respectively
 - E both numerical

The following information relates to Questions 5 to 8.

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



- 5** 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%
- 6** 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%
- 7** 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
- 8** 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 9 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

- 9** The range of meetings is:
A 22 **B** 23 **C** 24 **D** 43 **E** 44
- 10** The median number of meetings is:
A 6 **B** 7 **C** 7.5 **D** 8 **E** 9
- 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:

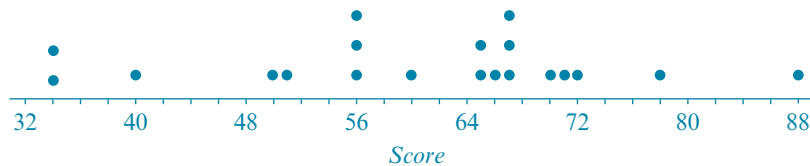
178.1 185.6 173.3 193.4 183.1 193.0

The mean and standard deviation are closest to:

- A** $\bar{x} = 184.4$; $s = 8.0$ **B** $\bar{x} = 184.4$; $s = 7.3$ **C** $\bar{x} = 182.5$; $s = 7.3$
D $\bar{x} = 182.5$; $s = 8.0$ **E** $\bar{x} = 183.1$; $s = 7.3$

The following information relates to Questions 13 and 14.

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



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

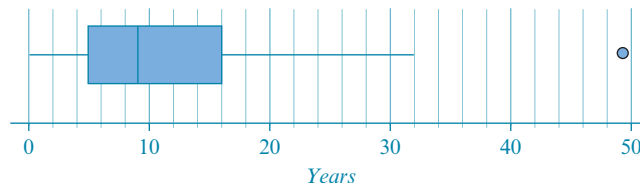
- A** 1 **B** 2 **C** 3 **D** 4 **E** 5

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

- A** 50% **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** 15 **B** 20 **C** 25 **D** 30 **E** 50

- 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 16 years is closest to:

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

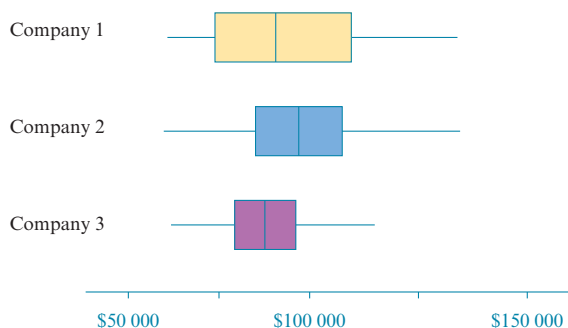
The following information relates to Questions 19 and 20.

The distribution of marks on a certain examination is approximately symmetric and bell shaped, with a mean of 65 and a standard deviation of 8.

- 19** Approximately what percentage of students score between 49 and 81?
A 50% **B** 68% **C** 95% **D** 99.7% **E** 100%
- 20** If 99.7% of students score between a and b marks, then possible values of a and b are:
A $a = 53, b = 77$ **B** $a = 57, b = 73$ **C** $a = 49, b = 81$
D $a = 41, b = 89$ **E** $a = 0, b = 100$

The following information relates to Questions 21 to 23.

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



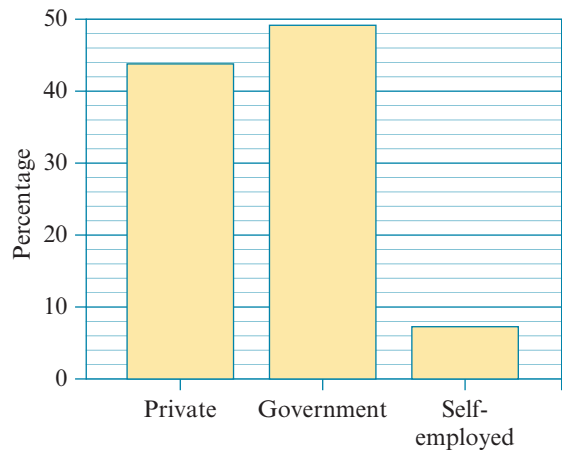
- 21** The company with the lowest median wage is:
A company 1 **B** company 2 **C** company 3
D company 1 and company 2 **E** company 2 and company 3
- 22** The company with the largest general spread (IQR) in wages is:
A company 1 **B** company 2 **C** company 3
D company 1 and company 2 **E** company 2 and company 3
- 23** 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 the 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 in company 2.

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 bar chart, shown opposite, 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?
c Given that 80 people were included in this study, how many identified as being employed in private companies?

- 3** A researcher asked a group of 30 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

- a** Using class intervals of width 5, construct a histogram of this data.
b Describe the shape of the histogram.
c Within which interval is the median number of cigarettes smoked?

- 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 the times taken.
b Make a stem plot of the times taken.
c Use this stem plot to find the median and quartiles for the time taken.
d Are there any outliers in this data? Justify your reasoning.

- 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.

- 6 A group of students was asked to record the number of SMS messages that they sent in one 24-hour period. Use the following five-number summary to construct a boxplot.

Min = 0, $Q_1 = 3$, $M = 5$, $Q_3 = 12$, Max = 24

- 7 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?

Written-response questions

- 1 A group of 500 people was asked to describe their general health by choosing one of the following responses: very healthy, pretty healthy, a little healthy, very unhealthy. The data is summarised in the following frequency table:

General health	Frequency	
	Number	%
very healthy	255	51.0
pretty healthy		35.6
a little unhealthy		
very unhealthy	29	
Total	500	100.0

- a What type of data has been collected: nominal, ordinal, discrete or continuous?
b Complete the frequency table and the percentage frequency table.
c Construct a percentage bar chart from the table.
d Write a report summarising the responses, quoting appropriate percentages to support your conclusion.
- 2 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 a dot plot of divorce rates.
 - c** What shape is the distribution of divorce rates?
 - d** What percentage of the 19 countries have divorce rates greater than 30%?
 - e** Calculate the mean and median of the distribution of divorce rates.
 - f** 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%?
- 3** 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?
 - b** Use your calculator to determine the following summary statistics for the *time* taken (rounded to two decimal places): \bar{x} , s , Min, Q_1 , M , Q_3 , 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: Min = 55, $Q_1 = 65$, $M = 70$, $Q_3 = 89$, 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).