

CHAPTER 1

Univariate data

eBookplus

DIGITAL DOC
doc-9523
10 Quick Questions

CHAPTER CONTENTS

- 1A Categorical data
- 1B Numerical data
- 1C Cumulative data
- 1D Measures of central tendency
- 1E Measures of variability
- 1F Stem-and-leaf plots
- 1G Boxplots
- 1H Comparing sets of data

1A Categorical data

Types of data

All data can be divided into two major groups:

- categorical data (also called qualitative data)
- numerical data (also called quantitative data)

Categorical data can be placed into 'categories'.

These categories can then be described as:

- nominal
- ordinal.

Numerical data are in the form of numbers and can be:

- discrete
- continuous.



Data — a valuable commodity

Data			
Categorical		Numerical	
Nominal	Ordinal	Discrete	Continuous
Need <i>sub-groups</i> (names) to complete the description, such as hair colour: blond, brown and so on.	Need a <i>ranking</i> to order the description, such as achievement levels: very high, high, satisfactory and so on.	<i>Counted</i> in exact values, such as goals scored in a football match, shoe size and so on. Values are often, but not always, whole numbers.	<i>Measured</i> in a continuous decimal scale, such as mass, time, length and so on.

WORKED EXAMPLE 1

Classify each of the following data using two types from the following descriptive words: categorical, numerical, nominal, ordinal, discrete, continuous.

- a The number of students absent from school.
- b The types of vehicles using a certain road.
- c The various pizza sizes available at a local takeaway.
- d The room temperature at various times during the day.





THINK

- a
 - 1 Determine whether the data are categorical or numerical.
 - 2 Determine whether the data are discrete or continuous.
- b
 - 1 Determine whether the data are categorical or numerical.
 - 2 Determine whether the data are nominal or ordinal.
- c
 - 1 Determine whether the data are categorical or numerical.
 - 2 Determine whether the data are nominal or ordinal.
- d
 - 1 Determine whether the data are categorical or numerical.
 - 2 Determine whether the data are discrete or continuous.

WRITE

- a The data are numerical as absences are represented by a number.
The data are discrete as the number of absences can be counted and are exact values.
- b The data are categorical as the types of vehicle need to be placed in categories.
There is no ranking or order involved, so the data are nominal.
- c The data are categorical as the pizza sizes do not represent a number, but a category.
The pizza sizes can be ranked from small to family, so the data are ordinal.
- d The temperature in a room is represented by a number, so the data are numerical.
Temperature can any value, and measurement is involved, so the data are continuous.

Working with categorical data

Once data have been collected, it is important to be able to summarise and display the data. For this purpose, two techniques are used:

- Frequency tables — these display the number of times each observation occurs
- Graphs — typically bar graphs and dot plots are used for categorical data

It is important to understand the terms used when analysing statistical data.

- The *frequency* of any observation is the number of times that that observation occurs and is given by the height of its column in a bar chart.
- The *relative frequency* of any observation is its frequency as a fraction of the total number of data entries.
- The *percentage frequency* is the relative frequency expressed as a percentage.

WORKED EXAMPLE 2

As part of a survey, a group of 30 teachers was asked to respond to the statement: ‘There is essentially no difference between the reasoning patterns used by girls and boys’. The teachers were asked to respond by writing T if they thought that the statement was true, F if they thought that the statement was false and U if they were unsure. The results were collated as follows.

T	F	F	F	T	F
T	U	T	F	T	U
U	F	T	F	T	T
T	U	U	F	T	F
F	F	U	T	U	T

- a Summarise the results using a frequency distribution table.
- b Represent the data by using a bar chart.
- c Find the frequency of teachers who thought that the statement was true.
- d Find the relative frequency of teachers who thought that the statement was true.
- e Find the percentage frequency of teachers who thought that the statement was true.

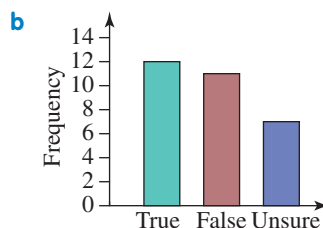
THINK

- a
 - 1 Rule up the table. Include a column for the type of response (category), a wide column for tallying the data and a column for frequency.
 - 2 Tally the data into a frequency distribution table.
 - 3 Fill in the frequency column by counting the number of tallies in each category.

WRITE/DRAW

Category	Tally	Freq.
T		12
F		11
U		7

- b** Draw the bar chart. There were 12 teachers who thought that the statement was true, 11 who thought that the statement was false and 7 who were unsure.



- c** The frequency of teachers who thought that the statement was true is simply the number that thought it was true; that is, the number of T entries in the frequency distribution table.

- c** Twelve teachers thought that the statement was true.

- d** The relative frequency of teachers who thought that the statement was true is the fraction of those who thought it was true out of the 30 teachers participating in the survey.

d $\frac{12}{30} = \frac{2}{5}$

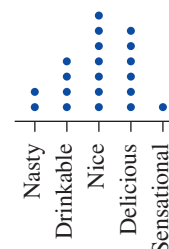
Two-fifths of the teachers thought that the statement was true.

- e** The percentage frequency is the relative frequency expressed as a percentage, so multiply the fraction by 100%.

e $\frac{2}{5} \times 100\% = 40\%$

Dot plots

- A dot plot is an alternative to a frequency distribution table.
- A dot is recorded for every piece of data in the correct position above a horizontal line.
- Dot plots can be used with both categorical and numerical data.



WORKED EXAMPLE 3

A group of 20 students was asked their reading preference. The survey revealed the following results.

comic	novel	newspaper	novel	newspaper
magazine	magazine	newspaper	novel	other
magazine	magazine	magazine	newspaper	comic
newspaper	other	magazine	newspaper	newspaper

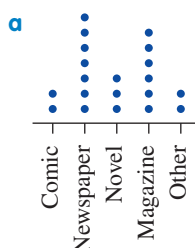
- a** Represent the data using a dot plot.
b What type of data are represented by the graph?

THINK

- a** 1 First, rule a horizontal line with each of the category names under it.
 2 Go through the responses piece by piece, entering each as a single dot on the dot plot.
 3 Cross out each response as it is entered.

Note: Remember to keep the dots in neatly spaced vertical columns.

WRITE/DRAW



- b** The categories are qualitative rather than quantitative so the data must be categorical. Because there is no real order into which the category names should be put, the observations are nominal data.

- b** Nominal categorical data

Exercise 1A Categorical data

- 1 **WE1** State whether the data formed in each of the following situations would be categorical or numerical.
 - a The number of matches in each box is counted for a large sample of boxes.
 - b The sex of respondents to a questionnaire is recorded as either M or F.
 - c The lengths of 40 cod are recorded by a fisheries inspector.
 - d The occurrence of hot, warm, mild and cool weather for each day in January is recorded.
 - e Cinema critics are asked to judge a film by awarding it a rating from one star to five stars.
- 2 State whether the categorical data formed by each of the following situations are nominal or ordinal.
 - a Individual members of a group of people are asked to indicate their level of agreement with the statement: 'I support open euthanasia'. Each respondent must select one of the following responses: strongly agree, agree, unsure, disagree or strongly disagree.
 - b Students are asked to write down the name of their favourite rock band.
 - c The mode of transport that each student uses to get to school is recorded.
 - d Members of a book club rate a new book as either 'a great read', 'quite readable' or 'not recommended'.
 - e The achievement level of each student in a class is recorded as either 'excellent', 'good', 'fair' or 'poor'.
- 3 State whether the numerical data formed by each of the following situations are discrete or continuous.
 - a The heights of 60 tomato plants at a plant nursery
 - b The number of jellybeans in each of 50 packets
 - c The time taken for each student in a class of 6-year-olds to tie their own shoes
 - d The petrol consumption rate of a large sample of cars
 - e The IQ (intelligence quotient) of each student in a class
- 4 **MC** Which of the following provides the best description of ordinal data?
 - A Ordinal data are categorical data in which there is no definite order associated with the categories used.
 - B Ordinal data are categorical data in which there is a definite order associated with the categories used.
 - C Ordinal data are numerical data that are arranged in ascending order.
 - D Ordinal data are discrete data that are arranged in ascending order.
 - E Ordinal data are any data that have been generated by random processes.
- 5 **MC** The number of people who are using a particular bus service are counted over a 2-week period. The data formed by this survey would be an example of:
 - A categorical data
 - B nominal data
 - C ordinal data
 - D discrete data
 - E continuous data
- 6 The following frequency distribution table gives the number of days of each weather type for the month of January. Represent the information by using a bar chart.

Weather type	Hot	Warm	Mild	Cool
Frequency	8	15	5	3

- 7 An international cricket player keeps a record of the way in which he has been dismissed when batting. Represent the information by using a bar chart.

Type of dismissal	Caught	Bowled	L. B. W.	Stumped	Run out	Hit wicket
Frequency	35	40	15	3	7	1

- 8 **WE2** A sample of students was asked: 'If you were able to choose this year's VCE subjects again would you still choose General Mathematics?' Their responses were divided into the categories Yes, No and Unsure. The results are shown below.

Y	Y	Y	N	U	Y
Y	N	N	U	U	Y
N	N	U	U	Y	Y
Y	Y	U	N	Y	N
U	Y	N	U	Y	Y

eBookplus

DIGITAL DOC
doc-9524
Graph types

eBookplus

DIGITAL DOC
doc-9525
SkillsHEET 1.1
Converting a fraction
to a decimal or
a decimal to a
percentage

- a Construct a frequency distribution table for the data.
- b Represent the information by using a bar chart.
- c How many students thought that they would still choose to study General Mathematics?
- d What was the relative frequency of students who would still choose to study General Mathematics?
- e What was the percentage frequency of students who would still choose to study General Mathematics?
- f What type of data does this question illustrate?

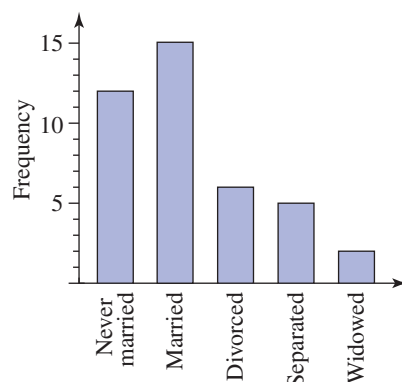
9 **MC** The following frequency table shows the music preference of a class of students.

Music preference	Metal	Classical	Rock	Jazz	Country
Frequency	6	2	15	5	8

What percentage of students prefer a type of music other than rock music?

- A 15%
- B 41.66%
- C 58.33%
- D 21%
- E None of these

10 The bar chart below shows the marital status of 40 respondents to a survey.



- a What was the most common marital status?
- b What was the frequency of responses marked 'married'?
- c What was the relative frequency of responses marked 'married'?
- d What was the percentage frequency of responses marked 'married'?
- e How many respondents were unmarried?
- f What percentage frequency of respondents were unmarried?

11 (You may like to attempt this question using a CAS calculator or computer spreadsheet package.)

Among other questions, a survey asked for the level of education of each participant. The data were organised into the following categories: X = (left school before Year 10), Y10 = (Year 10), Y12 = (Year 12), T = (technical college diploma or equivalent), U = (university degree).

Y10	Y12	Y12	U	U	T	Y12	Y10	U	T
Y12	Y12	Y12	U	T	T	Y10	Y12	U	T
U	Y10	T	X	U	T	U	Y12	Y12	T
Y12	Y12	T	U	Y10	T	Y12	Y12	X	Y12

- a Construct a frequency distribution table for the data.
- b Represent the information by using a bar chart.
- c How many participants achieved an education to Year 12 level?
- d How many participants achieved an education after Year 12 level?
- e What is the percentage frequency of the students who achieved an education after Year 12?
- f What type of data does this question illustrate?

12 **WE3** A shopkeeper keeps a record of how customers are paying for each transaction. The following is an extract showing a morning's trading.

credit card	account	account	cash	EFTPOS
cash	credit card	account	cash	cheque
cash	cash	credit card	credit card	cash
EFTPOS	account	cash	account	EFTPOS

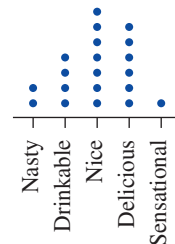
eBookplus

DIGITAL DOC
doc-9526
Dot plots

- a Construct a dot plot of the data.
- b How many transactions were conducted by cash?
- c What was the percentage frequency of transactions that were made by cash?
- d What type of data does this question illustrate?

13 **MC** This dot plot was drawn by a researcher who was asking people their opinions on the taste of a new soft drink. How many people participated in the survey?

- A 5 B 15 C 20 D 30 E 40



14 The audience of a talent quest show are asked to participate in the judging by rating each performance as either great, good, fair or poor. The following are the audience responses for a particular performer.

good great good good fair good
 great good great good good good
 good good fair fair great good
 good good great good poor great
 great fair good great good poor

- a Construct a dot plot of the data.
- b What was the opinion of the majority of the audience about the performance?
- c The overall score for each performer is calculated by awarding 4 points for each audience member who says that the performance was 'great', 3 points for each 'good', 2 points for each 'fair' and 1 point for each 'poor'. Find the score of this performer.

1B Numerical data

Scores

The remainder of this chapter is concerned with numerical data. With numerical data, each observation or data point is known as a *score*.

Grouping data

Numerical data may be presented as either *grouped* or *ungrouped* data.

The following table shows the number of cinema visits during a month by each of 20 students. It is an example of ungrouped data because each score is a separate group in itself.

Number of visits	0	1	2	3	4
Frequency	6	7	4	2	1

When there is a large amount of data or if the data are spread over a wide range it is useful to group the scores into *groups* or *classes*. The following table provides an example of grouped data. It shows the number of passengers on each of 20 bus trips.

Number of passengers	5–9	10–14	15–19	20–24	25–29
Frequency	1	6	8	4	1

When you are making the decision to summarise raw data by grouping it on a frequency distribution table, the choice of class size (group size) is important. As a general rule try to choose a class size so that 5 to 10 groups are formed. The choice of data labels is also important.

Consider the following data giving the number of nails in each of a sample of 40 boxes.

130 122 118 139 126 128 119 124 122 123
 132 138 129 139 116 123 126 128 131 142
 137 134 126 129 127 118 130 132 134 132
 137 124 134 134 120 137 141 118 125 129

Let's say that we wished to represent the data on a frequency distribution table. It would be pointless trying to tally the data as individual (ungrouped) scores as most scores would occur only once or twice

in the whole distribution. It makes good sense to group the scores. A class size of 5 would be most appropriate as this leads to the formation of 6 classes. Any of the following tally tables would be a possible (and correct) device for summarising the data.

Number of nails	115–120	120–125	125–130	130–135	135–140	140–145
Tally						
Frequency						

This method could get confusing for the person tallying because of uncertainty about where to place a data point if it fell on the boundary of two groups. In which group, for example, would you place the score 120?

Number of nails	115–119	120–124	125–129	130–134	135–139	140–144
Tally						
Frequency						

There is no ambiguity with this table and it is quite acceptable for this discrete data but it would not work when the data were continuous and numbers like 119.5 were possible. In any case this form of table leads to a situation which is difficult to graph. (We look at this later in this section.)

Number of nails	115–	120–	125–	130–	135–	140–
Tally						
Frequency						

This is probably the best frequency table for these data. The notation ‘115–’ should be read as ‘the group containing 115 and all the data right up to the start of the next group’. The three tables convey the same information but the last method will save possible problems when summarising and analysing the data.

The table, after tallying, would look like this:

Number of nails	115–	120–	125–	130–	135–	140–
Tally					I	
Frequency	5	7	10	10	6	2

Histograms and polygons

Numerical data can be graphed using histograms and polygons.

A *histogram* is similar to a bar chart but has the following essential features:

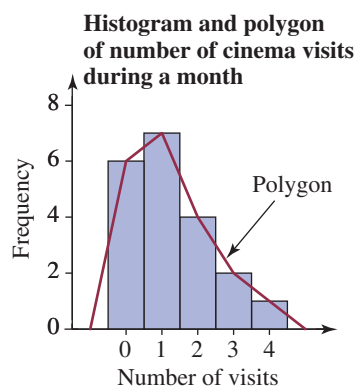
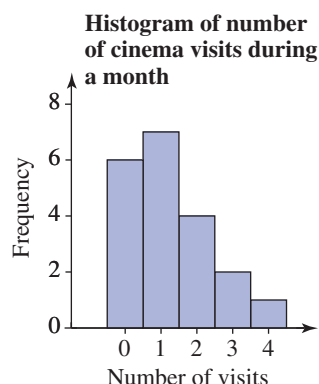
- Gaps are *never* left between the columns.
- If the chart is coloured or shaded then it is usually done all in the one colour. (The columns are essentially all representing different levels of the same thing.)
- Frequency is always plotted on the vertical axis.
- For *ungrouped* data the horizontal scale is marked so that the data labels appear *under the centre* of each column.
- For *grouped* data the horizontal scale is marked so that the end points of each class appear *under the edges of each column*.

An alternative method of displaying frequency distributions is by using a *polygon* (also called a *frequency polygon*).

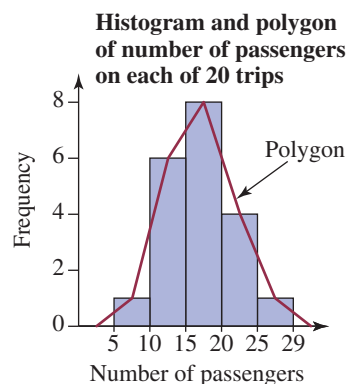
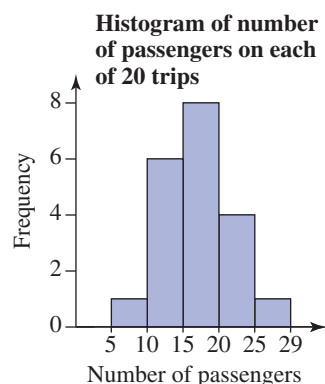
- A polygon is a line graph which can be drawn by joining the centres of the tops of each column of the histogram.
- The polygon starts and finishes on the horizontal axis a half column space from the group boundary of the first and last groups.

Histograms and polygons have been drawn below for the ungrouped cinema-visit data and the grouped bus-passenger data introduced previously in this section.

Number of visits	Frequency
0	6
1	7
2	4
3	2
4	1



Number of passengers	Frequency
5–9	1
10–14	6
15–19	8
20–24	4
25–29	1



- *Note:* The horizontal scale is marked under the *centre* of each column for the ungrouped cinema-visit data and at the *edges* of the columns for the grouped bus-passenger data.
- In the case of the bus-passenger data some liberty has been taken in adjusting the end points of each class to ensure that the horizontal scale is continuous. This had to be done so that no gaps would appear in between columns.

Histograms and polygons give an indication of the distribution of data: Are the data all crammed closely together or well spread? Are most of the data to one side of the distribution or the other?

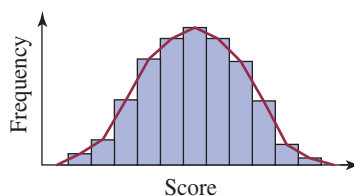
We associate certain names with different shapes of distribution. These include:

- *normal*
- *skewed*
- *bimodal*
- *spread*
- *clustered*.

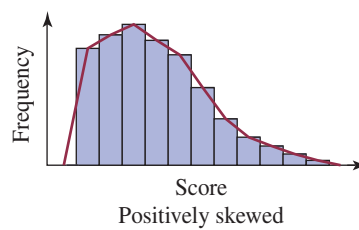
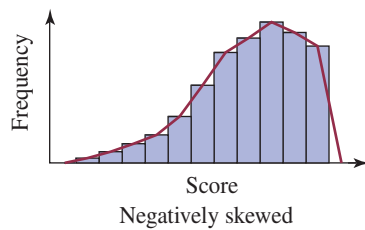
Consider the histograms and polygons below.

Normally distributed data

The most common score is located at the centre of the distribution. The further a score is away from this middle score, the less frequently it occurs. The distribution is symmetric around the central score.



Skewed data



The most common score is located towards the one end of the distribution. If the lower scores are spread over a greater range (tail on the left-hand side), the data are said to be *negatively skewed*. If the higher scores are spread over a greater range (tail on the right-hand side), then the data are said to be *positively skewed*.

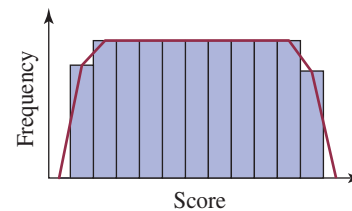
Bimodal data

There is more than one score that is most frequent.



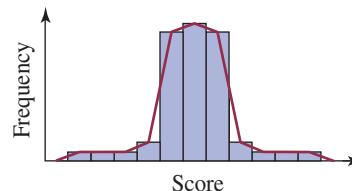
Spread data

The data are spread over a wide range.



Clustered data

Most of the data are confined to a small range.



WORKED EXAMPLE 4



The following data show the number of siblings of each of the 30 students in a particular class.

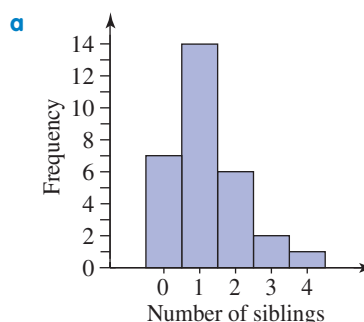
Number of siblings	0	1	2	3	4
Frequency	7	14	6	2	1

- Draw a histogram of the data.
- What was the frequency of the students with 2 siblings?
- What was the relative frequency of the students with 2 siblings?
- What was the percentage frequency of the students with 2 siblings?

THINK

- To draw a histogram:
 - Place the frequency on the vertical axis.
 - Place the number of siblings on the horizontal axis.
 - Draw a bar graph with no gaps between the adjacent columns; since these are ungrouped data, the data labels (that is, number of siblings) should appear under the centre of each column.

WRITE/DRAW





- b** Read the frequency of 2 siblings from the table.
- c** To find the relative frequency, express the number of students with 2 siblings as a fraction of the total number of students in the class.
- d** To find the percentage frequency, express the relative frequency as a percentage by multiplying it by 100%.

- b** Six students in the class had 2 siblings each.
- c** Relative frequency of the students with 2 siblings was $\frac{6}{30} = \frac{1}{5}$.
- d** $\frac{1}{5} \times 100\% = 20\%$
So 20% of all students in the class had 2 siblings each.

WORKED EXAMPLE 5

The following data give the weights (in kg) of a sample of 25 Atlantic salmon selected from a holding pen at a fish farm.

10.2	12.6	10.4	9.8	12.2
8.7	10.4	11.3	12.2	14.1
10.8	10.7	9.5	13.4	8.8
10.0	12.1	11.4	11.7	10.4
11.0	10.4	10.9	9.6	8.8

eBookplus

TUTORIAL
eles-1337
Worked example 5

- a** Represent the data on a frequency distribution table.
- b** Draw a histogram of the data.
- c** Add a polygon to the histogram.
- d** What word could you use to describe the pattern of the distribution of the data? Choose one of the following: normal, positively skewed, negatively skewed, bimodal, clustered or spread.

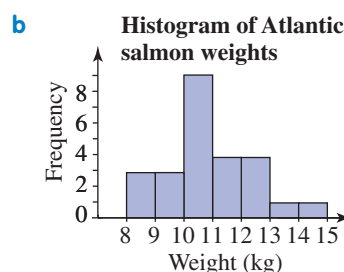
THINK

- a** Because the scores are very diverse it is necessary to group them before attempting to tally the data. A group size of 1 kg would seem most appropriate. The data labels on the frequency distribution table would be best written as 8–, 9–, 10–, rather than 8–9, 9–10, 10–11, to avoid any ambiguity when tallying.

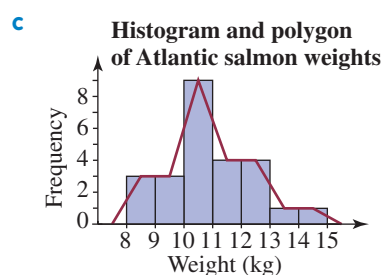
WRITE/DRAW

Weight (kg)	Tally	Frequency
8–		3
9–		3
10–		9
11–		4
12–		4
13–		1
14–		1

- b** Frequency is always assigned to the vertical axis of a histogram. No gaps are left between adjacent columns. When a scale is placed on the horizontal axis the group boundaries are positioned at the edges of each column (as these are grouped data).



- c** The polygon can be formed by connecting the tops of each column in the histogram.



- d** The most common score is not in the centre; the higher scores are spread over a greater range.

- d** The data are positively skewed.

Exercise 1B Numerical data

- 1 **WE4** The following data show the number of registered cars normally kept at each of 30 households.

Number of cars	0	1	2	3	4	5
Frequency	2	8	11	6	2	1

- Draw a histogram of the data.
- What was the frequency of a household with 3 cars?
- What was the relative frequency of a household with 3 cars?
- What was the percentage frequency of a household with 3 cars?

- 2 The following table shows the number of jelly beans in each of 60 packets.

Number of jelly beans	48	49	50	51	52	53
Frequency	2	10	32	9	5	2

- Draw a histogram of the data.
- What was the frequency of a packet with 50 beans?
- What was the relative frequency of a packet with 50 beans?
- What was the percentage frequency of a packet with 50 beans?

- 3  The following data give the number of hours of study completed by 30 students on a certain night.

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

- Represent the data using a dot plot.
- With the aid of the dot plot, rearrange the data into a frequency distribution table.
- Draw a histogram of the data.
- Add a polygon to the histogram.
- What was the most common number of hours of study?
- What was the frequency of 2 hours of study?
- What was the relative frequency of 2 hours of study?
- What was the percentage frequency of 2 hours of study?

- 4 The following frequency table gives the number of oysters of different lengths from a tray in a marine farm.

Length (cm)	4–	5–	6–	7–	8–	9–
Frequency	6	10	60	58	8	4

- Draw a histogram of the data.
- Find the frequency, the relative frequency and the percentage frequency of oysters of size 5–6 cm.

- 5 The following frequency table gives the results of testing the lives of 200 torch batteries.

Lifetime (hours)	20–	25–	30–	35–	40–	45–
Frequency	6	25	70	61	30	8

- Draw a histogram of the data.
- Find the frequency, the relative frequency and the percentage frequency of batteries that lasted between 30 and 35 hours.

- 6  The following frequency table gives the number of crayfish that a fisherman caught on each of 78 trips that he made during a season.

Number of crayfish	30–39	40–49	50–59	60–69	70–79	80–89	90–99
Frequency	12	20	17	15	7	6	1

- Draw a histogram of the data.
- Add a polygon to the histogram.
- What word could you use to describe the pattern of the distribution of the data? Choose one of the following: normally distributed, positively skewed, negatively skewed, bimodal, clustered or spread.

eBookplus

DIGITAL DOC
doc-9527
One-variable statistics

eBookplus

DIGITAL DOC
doc-9528
One-variable
statistics —
grouped

- 7 The following data give the length of the forearm (in cm) of 30 middleweight boxers.

24.0 26.5 31.4 30.8 25.7 26.9 27.2 29.9 25.2 26.3
 28.4 26.3 29.2 30.5 26.0 27.2 26.0 24.9 25.8 24.7
 28.3 26.3 25.7 26.9 27.8 25.7 25.0 29.3 25.8 28.0

- a Construct a frequency distribution table for the data. Use a class size of 1 cm.
 b Draw a histogram of the data. (You might like to attempt this using a CAS calculator.)
 c What word could you use to describe the pattern of the distribution of the data?
- 8 **WE5** The following data give the times (in seconds) taken for athletes to complete a 100 m sprint.

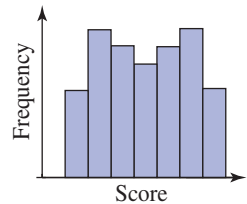
12.2 12.0 11.9 12.0 12.6 11.7 11.4 11.0 10.9 11.7
 11.2 11.8 12.2 12.0 12.7 12.9 11.3 11.2 12.8 12.4
 11.7 10.8 13.3 11.7 11.6 11.7 12.2 12.7 13.0 12.2

- a Construct a frequency distribution table for the data. Use a class size of 0.5 seconds.
 b Draw a histogram of the data. (You might like to attempt this using a CAS calculator.)
 c Add a polygon to the data.
 d What word could you use to describe the pattern of the distribution of the data?
- 9 The data below show the fat content (%) of 20 packs of pork bacon selected from a supermarket's shelves.

6 13 3 8 9
 15 18 13 4 25
 21 13 18 20 8
 9 16 15 18 12

- a Construct a frequency distribution table for the data using a class size of 4.
 b Construct a frequency distribution table for the data using a class size of 10.
 c Comment upon how changing the size of the group affects the appearance of the histogram.
- 10 **MC** What type of data is represented by the histogram at right?

- A Positively skewed data
 B Negatively skewed data
 C Clustered data
 D Spread data
 E Bimodal data



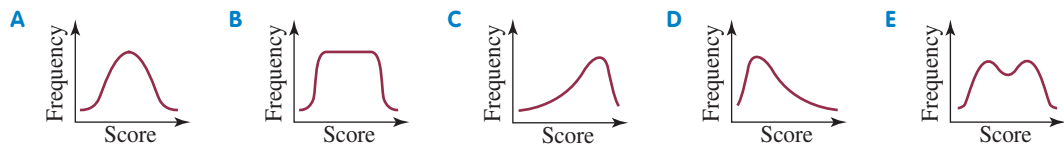
- 11 **MC** What type of data is represented by the frequency distribution table below?

Score	60–	70–	80–	90–	100–
Frequency	2	2	35	2	1

- A Positively skewed data B Negatively skewed data C Clustered data
 D Spread data E Bimodal data

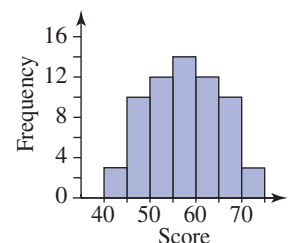
- 12 **MC** Match the following frequency distribution table with its likely polygon.

Score	55–	60–	70–	80–	90–	100–
Frequency	1	3	8	16	25	10



- 13 **MC** Which of the following statements is *untrue* of the data represented by the histogram at right?

- A The data appear to be normally distributed.
 B The class size is 5.
 C The class 55–60 has the greatest frequency.
 D The percentage frequency of the class 40–45 is 3%.
 E About 64 scores were involved in the construction of this histogram.



1C Cumulative data

Cumulative frequency

It is often useful to consider the number of data points that are less than or equal to a particular score. In such cases it is helpful to include a *cumulative frequency* column on the frequency distribution table.

The cumulative frequency is the number of records equal to and less than a particular score.

- If all the data were sorted in order of size, the cumulative frequency would give a ‘running total’ of the number of observations up to each score.

Consider the following data that show the heights (in cm) of 40 girls who are competing in trials to form a basketball squad.

181	191	185	175	192	186	188	182	179	172
186	188	193	198	182	175	176	188	180	191
185	186	183	180	179	175	180	188	190	193
172	179	188	183	186	188	182	193	194	181

A frequency distribution table including a cumulative column could be drawn as follows:

Height (cm)	Frequency	Cumulative frequency
170–	3	3
175–	6	9
180–	12	21
185–	10	31
190–	8	39
195–	1	40

The cumulative frequency column in this case records the number of girls who had a height as indicated by the particular group or those preceeding it. For example the figure 31 in the cumulative frequency column can be interpreted as: ‘There were 31 girls with a height of less than 190 cm. The figure 31 in this example was found by totalling $3 + 6 + 12 + 10$. Note that the final number in the cumulative frequency column should always equal the total number of scores.

Ogives

An ogive (also called a cumulative frequency polygon) is a line graph of the cumulative frequency results.

- An ogive is appropriate only for displaying grouped data.
- The graph is started on the horizontal axis at a point corresponding to the lowest possible score in the smallest group. In the case of the basketball squad data the graph will start at 170; that is, the first point on the graph is (170, 0).
- The ogive is then drawn by plotting the value of the cumulative frequency of each group against each group end point. For the basketball squad data the points which form the rest of the ogive will be (175, 3), (180, 9), (185, 21), (190, 31), (195, 39) and (200, 40).
- The ‘S’ shape of this ogive is typical of most sets of data.
- A percentage axis can be added by ruling a vertical line from the end point of the ogive to the horizontal axis. The end point is labelled 100% and then the axis is scaled from 0–100% appropriately.

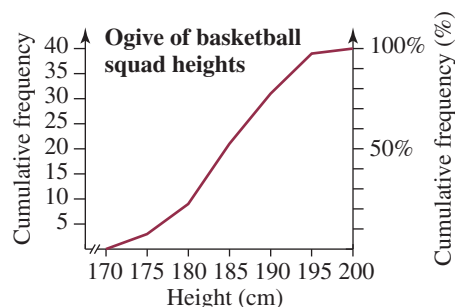


Figure A

A percentage axis is not an essential feature of an ogive but it will help to answer a lot of questions like some of the following:

1. How many girls had a height of less than 182 cm?
(Find 182 on the height axis, rule a vertical line to the ogive then horizontally to the frequency axis. (See line (a) in figure B.) Answer: About 15 girls would have a height of less than 182 cm.
 2. What percentage of girls had a height of less than 180 cm? (Find 180 on the height axis then rule a vertical line to the ogive then horizontally to the percentage axis. See line (b) in figure B.) Answer: About 22% of girls had a height of less than 180 cm.
 3. What percentage of girls had a height more than 180 cm? Answer: About 22% had a height of less than 180 cm so there must be 78% with a height of more than 180 cm.
- Note that when you are interpreting ogives, less than ($<$) and less than or equal to ($\leq$) make no difference to the approach that we take to solving a question or its answer. Example 1 above would have had exactly the same answer if the question had been: 'How many girls had a height of less than or equal to 182 cm?'

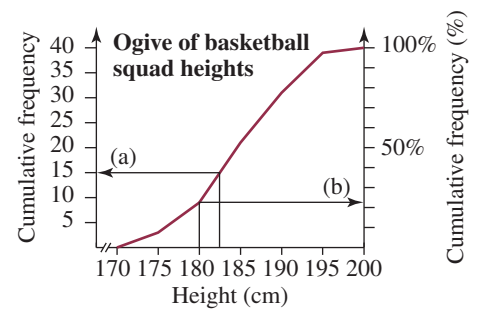


Figure B

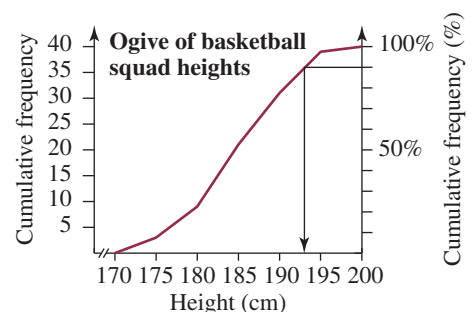


Percentiles

A percentile is the score below which a particular percentage of the distribution of data lies.

For example the 90th percentile is the score below which 90% of the data lies. In the case of the basketball squad data the 90th percentile could be found by finding 90% on the percentage axis, going horizontally to the ogive, then vertically down to the height axis.

The 90th percentile would be about 193 cm. This could be interpreted as: '90% of the girls would have a height of 193 cm or less'.



WORKED EXAMPLE 6

eBookplus

Forty sample pieces of rope are tested in an effort to determine their breaking strain. The maximum load that could be attached to each was recorded.

TUTORIAL
eles-1338
Worked example 6

Breaking strain (kg)	Frequency
40–	2
45–	6
50–	8
55–	10

Breaking strain (kg)	Frequency
60–	9
65–	4
70–	1

- Add a cumulative frequency column to the table.
- Represent the data using an ogive.
- What number of sample pieces broke under a strain of less than 52 kg?
- Find the 75th percentile and write a sentence to explain what it means.
- The manufacturer of the rope wishes to label the rope with an appropriate breaking strain. What should the rope be rated at if the manufacturer wants 90% of all ropes to be at least as strong as the labelled rate?

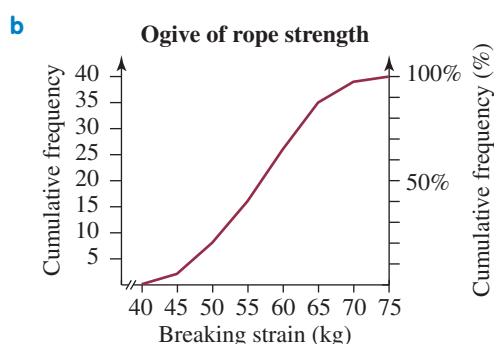
THINK

- The cumulative frequency column is a 'running total' of the amounts in the frequency column. The final entry in the cumulative frequency column should match the number of observations.

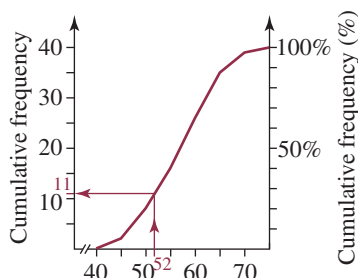
WRITE/DRAW

- | Breaking strain (kg) | Frequency | Cumulative frequency |
|----------------------|-----------|----------------------|
| 40– | 2 | 2 |
| 45– | 6 | 8 |
| 50– | 8 | 16 |
| 55– | 10 | 26 |
| 60– | 9 | 35 |
| 65– | 4 | 39 |
| 70– | 1 | 40 |

- The maximum height of the ogive will be 40 because there were 40 observations. The ogive will start at 40 kg on the horizontal axis. The next point will represent the end of the first group. By the end of the first group (that is, 45 kg) there had been 2 observations entered. So (45, 2) is the next point on the ogive. The last point is (75, 40). Join the points to complete the ogive. Draw a vertical line from the end point of the ogive to the horizontal axis. Label the top end point as 100%, then scale the axis appropriately.



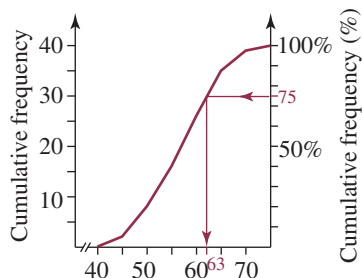
- Find 52 kg on the horizontal axis, go up to the ogive, then along to the frequency axis. Arrive at 11.



- About 11 pieces of rope broke under a strain of less than 52 kg.

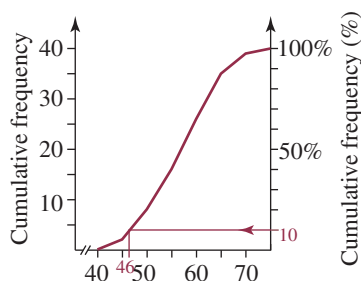


- d Find 75% on the percentage axis, go across to the ogive, then down to the horizontal axis.



- d The 75th percentile is 63 kg. That is, 75% of the sample pieces broke under a strain of 63 kg or less.

- e If 90% of the ropes are to withstand the strain then 10% of the ropes will break with this strain or less. So find the 10th percentile.



- e We see that 10% of the ropes will break under a strain of less than 46 kg. So if the rope is marketed as 46 kg breaking strain then 90% of the ropes will withstand the strain.

Find 10% on the percentage axis, go across to the ogive, then down to the horizontal axis.

Exercise 1C Cumulative data

eBookplus

DIGITAL DOC
doc-9529
One-variable
statistics

- 1 **WE6a,b** The frequency table below shows the lengths of 77 flathead caught during a fishing competition.

- a Copy the table and add a cumulative frequency column to it.
b Prepare an ogive of the data.

Length of fish (mm)	Frequency
300–	9
310–	15
320–	20
330–	12

Length of fish (mm)	Frequency
340–	8
350–	7
360–	4
370–	2

- 2 The following frequency table shows the times taken for 60 people involved in a psychology experiment to complete a simple manipulative puzzle.

Time taken (sec)	Frequency
6–	1
8–	4
10–	15
12–	18

Time taken (sec)	Frequency
14–	12
16–	8
18–	2

- a Copy the table and add a cumulative frequency column to it.
b Prepare an ogive of the data.

- 3 **WE6** The salaries of the 40 employees of a small manufacturing company are represented by the accompanying frequency table.

Salary (\$1000)	Frequency
15–	6
20–	12
25–	8
30–	7

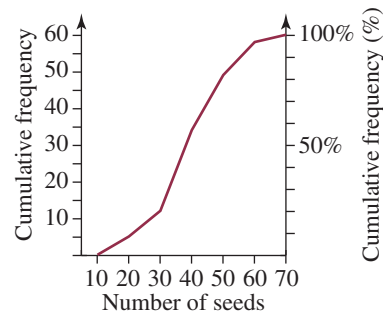
Salary (\$1000)	Frequency
35–	5
40–	1
45–	1

- a Copy the table and add a cumulative frequency column to it.
 - b Prepare an ogive of the data.
 - c How many employees are earning less than \$22 000?
 - d How many employees are earning less than \$31 000?
 - e Find the 75th percentile of the data and write a sentence explaining what it means.
 - f Find the 50th percentile of the data and write a sentence explaining what it means.
 - g Find the 25th percentile of the data and write a sentence explaining what it means.
 - h The management decides to award pay rises to its highest earning employees. The top 10% of employees will all get a pay rise. How much salary would an employee need before qualifying for a pay rise?
- 4 A manufacturer of surf clothing needs to know how many clothes of different sizes to produce. The manager organises a survey of young people which provides the following data:

Waist size (cm)	Frequency
70–	13
75–	28
80–	46
85–	30

Waist size (cm)	Frequency
90–	17
95–	8
100–	7
105–	1

- a Copy the table and add a cumulative frequency column to it.
 - b Prepare an ogive of the data.
 - c How many young people had a waist size of less than 82 cm?
 - d How many young people had a waist size of greater than 94 cm?
 - e Find the 90th percentile of the data and write a sentence explaining what it means.
 - f Find the 50th percentile of the data and write a sentence explaining what it means.
 - g The manager decides that production costs can be minimised by only making garments fitting sizes between 78 cm and 100 cm. What percentage of the population will not be catered for by this manufacturer?
- 5 A biologist who counts the number of seeds in each of 60 pumpkins presents his findings on the ogive at right.
- a How many pumpkins contained 30 or fewer seeds?
 - b How many pumpkins contained more than 50 seeds?
 - c What percentage of pumpkins had fewer than 45 seeds?
 - d What percentage of pumpkins had fewer than 20 seeds?
 - e Find the 90th percentile of the data and write a sentence explaining what it means.
 - f Find the 75th percentile of the data and write a sentence explaining what it means.
 - g Find the 50th percentile of the data and write a sentence explaining what it means.
 - h The worst 20% of pumpkins (those with the fewest seeds) are to be kept aside for further investigation. Find the maximum number of seeds for any pumpkin in this group.



- 6 A time trial is a race in which each competitor rides separately, racing 'against the clock'. The following are the times (in seconds) of 20 competitors in a 1-km cycling time trial.

75	72	68	78	75
68	77	80	85	82
73	75	82	90	92
75	73	72	70	83

- a Copy and complete the frequency distribution table overleaf.
- b Prepare an ogive of the data.

- c How many riders finished with a time of 82 seconds or better?
- d Find the 90th percentile and write a sentence which explains what it means.
- e The top 20% of riders are to be selected for a special training squad. What time would be needed to qualify for the squad?

Time	Tally	Frequency	Cumulative frequency
65–			
70–			
75–			
80–			
85–			
90–			

- 7 The following data, collected from a maternity hospital, give the birth weights (in kg) of 30 babies.

3.7 3.2 3.8 4.1 2.9 3.3 3.6 3.1 3.6 3.9
 4.2 2.5 2.7 3.9 3.6 3.2 3.0 2.9 3.4 3.0
 3.1 2.8 2.9 3.2 3.1 3.8 3.9 3.3 4.4 3.4

- a Copy and complete the frequency distribution table at right.
- b Prepare an ogive of the data.
- c What percentage of babies had a birth weight of 3.8 kg or less?
- d Find the 50th percentile and write a sentence that explains what it means.
- e Babies of weight 2.6 kg or less are given special attention by medical staff. What percentage of babies will be given special attention?

Weight	Tally	Frequency	Cumulative frequency
2.4–			
2.8–			
3.2–			
3.6–			
4.0–			
4.4–			

- 8 (You might like to attempt this question using a CAS calculator.) The number of people who are using a particular airline flight is recorded on a weekly basis for 20 weeks. The results are shown at right:

85 103 98 72 88
 86 92 108 112 84
 95 99 84 79 88
 93 87 79 81 80

- a Summarise the data using a frequency distribution table. Use a class size of 10. Include a cumulative frequency column.
 - b Prepare an ogive of the data.
 - c How many of the flights had fewer than 95 passengers?
 - d Find the 75th percentile and write a sentence which explains what it means.
 - e Find the 50th percentile and write a sentence which explains what it means.
 - f Find the 25th percentile and write a sentence which explains what it means.
 - g The management of the airline is considering using a smaller aircraft for the flight. One of the management team calculates that it would be more profitable to use a smaller plane if 20% of current flights have 80 or fewer passengers. Should a smaller aircraft be used?
- 9 (You might like to attempt this question using a CAS calculator.) Thirty samples of paint are exposed to high temperature until they begin to blister. The following data show the temperature (°C) at which each began to blister.

75 88 76 92 77 88 82 81 93 96
 81 88 75 79 80 81 77 88 92 79
 81 83 88 82 85 79 93 84 84 77

- a Represent the data on a frequency distribution table using a class size of 5 °C. Include a cumulative frequency column.
- b Prepare a histogram of the data.
- c Prepare an ogive of the data.
- d How many paint samples blistered at a temperature of 85 °C or less?
- e What percentage of the samples blistered at a temperature of 85 °C or less?
- f What percentage of the samples could withstand a temperature of 88 °C?
- g Find the 90th percentile and write a sentence explaining what it means.
- h The manufacturer wishes to guarantee the paint at high temperatures. What is the highest temperature at which the paint could be rated if the management wants at most 20% of the paint returned because it failed to meet its rating?

- 10 **MC** Which of the following is *untrue* of ogives?
- A Ogives display cumulative data.
 - B Ogives are a type of line graph.
 - C Ogives tell the number of scores that are of a particular value or less.
 - D An ogive cannot be interpreted without having the original experimental data at hand.
 - E An ogive can also be called a cumulative frequency polygon.
- 11 **MC** The 90th percentile of some data is 25. This means that:
- A 10% of the data are less than or equal to 25
 - B 10% of the data are greater than or equal to 25
 - C 90% of the data are less than or equal to 25
 - D 90% of the data are greater than or equal to 25
 - E 25% of the data are less than or equal to 90

eBookplus

DIGITAL DOC
doc-9530
WorkSHEET 1.1

1D Measures of central tendency

It was said earlier that one of the main tasks of the statistician is to summarise large volumes of data. It is useful to be able to find one score which is typical of a whole set of data, or a few figures which can describe its distribution.



Finding the mean, median and mode are three different methods of arriving at a score which is typical or central to a set of data. Mean, median and mode are often called *measures of central tendency*.

The mean

The most commonly used measure is the *mean*.

The mean is the average score in the set.

- To find the mean all the scores are added, then the total is divided by the number of scores. The symbol $\bar{x}$ is commonly used to denote the mean.
- The operation of finding the mean could be written as a formula: $\bar{x} = \frac{\sum x_i}{n}$.
- The symbol Σ is the Greek capital letter sigma. In maths it is used to signify 'the sum of'.

So this formula $\bar{x} = \frac{\sum x_i}{n}$ could be read as: 'The mean equals the sum of all the scores (x) divided by the number of scores (n)'.

The following data give the number of pets kept in each of 10 different households:

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

We could find the mean as follows:

$$\text{Mean} = \bar{x} = \frac{\sum x_i}{n} = \frac{3+5+4+4+2+3+0+1+4+5}{10} = 3.1$$

The median

The *median* of a set of scores is the middle score when the data are arranged in order of size.

- The position of the median can be found using the formula:

$$\text{median position} = \frac{n+1}{2} \text{ th score}$$

- For an odd number of scores, the median will be one of the scores in the distribution.
- For an even number of scores, the median will occur halfway between two scores

There were 10 scores in the household pets data, so the median is the $\frac{10+1}{2}$ th score; that is, the 5.5th score or halfway between the 5th and 6th scores when they are arranged in order of size. To find the median for the household pets data we must first arrange them in order of size:

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

The 5th score is 3. The 6th score is 4. The median is halfway between them. This number can be found by averaging the two.

$$\begin{aligned}\text{Median} &= \frac{3+4}{2} \\ &= 3.5\end{aligned}$$

The mode

The *mode* of a group of scores is the score that occurs most often.

- The mode is the score with the highest frequency.

In the case of the household pets data, a score of 4 occurs three times, making it the most frequently occurring score. It is therefore the mode of the data. In some cases there will be two or more scores which occur equally 'most often'. In such cases all of them are modes. (Don't attempt to average them.)

WORKED EXAMPLE 7

The following data give the number of hours spent on homework by 8 students:

2, 2, 3, 0, 1, 1, 5, 1.

- a** Determine the mean of the data.
- b** Determine the median of the data.
- c** Determine the mode of the data.

THINK

- a** The mean is the average of the scores, so add them and divide by 8 because there are 8 numbers in the distribution.
- b**
 - 1 First arrange the numbers in order of size.
 - 2 The position of the median is given by the formula: $\frac{n+1}{2}$ th score.
 - 3 When the median is halfway between two scores it can be found by averaging them.
- c** The mode is the score that occurs most often.

WRITE

$$\begin{aligned}\text{a Mean} &= \frac{2+2+3+0+1+1+5+1}{8} \\ &= 1.875\end{aligned}$$

$$\text{b } 0, 1, 1, 1, 2, 2, 3, 5$$

$$\begin{aligned}\text{Median} &= \frac{8+1}{2} \text{ th score} \\ &= 4.5 \text{th score} \\ &\text{that is, halfway between the 4th and 5th scores.}\end{aligned}$$

The 4th score is 1. The 5th score is 2.

$$\begin{aligned}\text{Median} &= \frac{1+2}{2} \\ &= 1.5\end{aligned}$$

$$\text{c Mode} = 1$$

If the data have been presented as an ungrouped frequency table (that is, each group is just one number) then our techniques have to be adapted a little in order to find the mean, median and mode.

WORKED EXAMPLE 8



The table shows the number of cinema visits made by each of 20 students.

Find:

- a the mean of the data
- b the median of the data
- c the mode of the data.

Number of visits	0	1	2	3	4
Frequency	6	7	4	2	1

THINK

First add extra columns to the frequency distribution table: $f \times x$ column and cumulative frequency column. The $f \times x$ column can be used to help total the scores by subtotalling the scores in each class. Entries in this column can be found by multiplying each score by its frequency. The total of this column is the total of all the scores.

WRITE

No. of visits	Freq.	$f \times x$	Cumul. freq.
0	6	0	6
1	7	7	13
2	4	8	17
3	2	6	19
4	1	4	20
Total		25	

- a To find the mean of the data, divide the total of all scores by the number of students.
- a Mean = $25 \div 20$
= 1.25
- b 1 Find the position of the median, which is the $\frac{n+1}{2}$ -th score.
- b The median is the $\frac{20+1}{2}$ th score —
which is the 10.5th score — halfway between the 10th and 11th scores.
The median is 1.
- 2 The scores are already arranged in order on the frequency distribution table. The cumulative frequency column tells us that the 6th score was a 0. Then all the scores up to the 13th score were all the number 1. We are interested in only the 10th and 11th scores which both must be 1. So the median is 1.
- c The mode is the score with the highest frequency. This is easy to find when the data are presented as a frequency distribution table. For the cinema data the score of 1 occurred 7 times and no other score occurred as often.
- c The mode is 1.

When working with data that have been grouped in a frequency distribution table, we have to take into account several extra considerations.

- When completing an $f \times x$ column we are unable to distinguish the original scores as they have been lost due to the grouping. All we can do is assume that all the scores were equal to the centre or *midpoint* of the group. It is useful to add a midpoint column to the data. Note that this may yield a slightly inaccurate mean.
- It becomes very difficult to find the exact position of the median with such data. One way it could be located is by finding the 50th percentile from an ogive of the data. (The median is approximately equal to the 50th percentile.) Or, rather than locating exactly the median, the *median class* (the group that contains the median) may be all that is required. This could be read easily from the frequency distribution table.
- The *modal class* (the class that contains the median) is a more appropriate statistic to use than the mode.

WORKED EXAMPLE 9

eBookplus

The frequency table below shows the area (in m^2) of the 23 blocks in a suburban subdivision.

- Find the mean block size.
- Find the median class for block size.
- Find the modal class for block size.

TUTORIAL
eles-1339
Worked example 9

Area (m^2)	520–	540–	560–	580–	600–	620–	640–
Frequency	3	5	7	3	2	2	1

THINK

Add 3 extra columns to the frequency distribution table: a midpoint column, an $f \times x$ column and a cumulative frequency column. Fill in the 3 columns as follows:

- To complete the midpoint column: The first group is for blocks that are of size 520–540 m^2 . So the midpoint is 530 m^2 . The second group is for blocks that are of size 540–560 m^2 . So the midpoint is 550 m^2 , etc.
- To complete the $f \times x$ column: There are 3 blocks of size about 530 m^2 so their total area is $3 \times 530 = 1590 \text{ m}^2$. There are 5 blocks of size about 550 m^2 so their total area is $5 \times 550 = 2750 \text{ m}^2$ and so on.
- To complete the cumulative frequency column: Copy the frequency for the first group (3) into the cf column. To find the cf for every other group, add its frequency to the preceding cf value. For example, for the second group $\text{cf} = 3 + 5 = 8$, for the third group $\text{cf} = 8 + 7 = 15$, and so on.

WRITE

Area	f	Mid pt.	$f \times x$	cf
520–	3	530	1590	3
540–	5	550	2750	8
560–	7	570	3990	15
580–	3	590	1770	18
600–	2	610	1220	20
620–	2	630	1260	22
640–	1	650	650	23
	23		13 230	

- To find the mean, total the $f \times x$ column, then divide by 23 because there were 23 blocks in the subdivision.

$$\begin{aligned} \text{a Mean} &= 13\,230 \div 23 \\ &= 575.22 \text{ m}^2 \end{aligned}$$

- Find the position of the median, using the formula:
Median position = $\frac{n+1}{2}$ th score.

- The median is the $\frac{23+1}{2}$ th score — that is, the 12th score.

- Use the cumulative frequency column to determine the group that contains the 12th score. The group 540–560 contains the 8th score. The group 560–580 contains the 9th score to the 15th score. So the 12th score must be in the group 560–580.

The median class is 560–580 m^2 .

- The modal class is the one that occurs with the greatest frequency. The group 560–580 has the highest frequency (7).

- The modal class is 560–580 m^2 .

The effect of extreme values

An *extreme value* (or *outlier*) is a score which is considerably different from the rest of the data in the set. The presence of extreme values may affect the representative nature of any statistics calculated.

Consider again the number of household pets data that were introduced earlier in this section: 3, 5, 4, 4, 2, 3, 0, 1, 4, 5.

We have already determined the mean, median and mode:

Mean = 3.1

Median = 3.5

Mode = 4.

Let's include an extra household in this survey — the household of dear old Mrs Pinebottom and her 99 cats! How would this extra piece of data affect the mean, median and mode?

There would be 11 pieces of data: 3, 5, 4, 4, 2, 3, 0, 1, 4, 5, 99.

We would find that:

Mean = 11.82 Median = 4 Mode = 4. (Check these calculations for yourself.)

The inclusion of an extreme value has:

- dramatically increased the mean of the data
- marginally increased the value of the median
- left the mode unchanged.

The important point to learn from this illustration is that when a set of data includes extreme values then the mean may not be truly representative of typical scores in the set. In the case of the household pets data none of the households had a number of pets near 11.82. Under such circumstances the median or mode may be a more useful way of relating the notion of a typical score.

Exercise 1D Measures of central tendency

- 1 **WE7** Find the mean, median and mode of each of the following sets of data.

a 6, 8, 4, 7, 6, 7, 6

b 102, 99, 106, 108, 101, 103, 102, 105, 102, 101

c 5.6, 5.2, 5.4, 5.3, 5.8, 5.4, 5.3, 5.4

d 25, 27, 29, 27, 26, 28, 29, 28, 27, 28

- 2 **WE8** The number of daily errors that a machine makes is recorded over a 50-day period. The following frequency table is obtained.

Number of errors per day	0	1	2	3	4	5
Frequency	9	18	13	6	3	1

- a Find the mean number of errors that the machine makes in a day.
b Find the median number of errors that the machine makes in a day.
c Find the modal number of errors that the machine makes in a day.
- 3 **WE9** The number of days off work due to sickness is recorded for each worker in a small business over the course of a year. The following frequency distribution table is compiled.

Days of sickness	0–4	5–9	10–14	15–19	20–24	25–29	30–34
Frequency	10	10	7	6	5	2	1

- a Find the mean number of days of sickness for an employee.
b Find the median class for the number of days of sickness for an employee.
c Find the modal class for the number of days of sickness for an employee.
- 4 The depth of snow at a ski resort is measured every day over the course of a ski season. The following frequency distribution table is compiled.

Depth (cm)	Frequency
0–	8
50–	9
100–	12
150–	15

Depth (cm)	Frequency
200–	6
250–	4
300–	2
350–	2

- a Find the mean depth of snow during the ski season.
b Find the median class for snow depth.
c Find the modal class for snow depth.

eBookplus

DIGITAL DOC
doc-9527
One-variable
statistics

- 5 **MC** A correct description of how to find the mean of a group of scores is:
- A Find the most common score
 - B Total the scores, then divide by the number of scores
 - C Count the scores, add 1, then divide by 2
 - D Find the middle score when the scores are arranged in order of size
 - E Average greatest and least scores
- 6 **MC** There are seven scores in a set of data. The scores have been arranged in order of size. Which of the following will be true of the median?
- A The position of the median is the 3.5th score.
 - B The position of the median is the 4th score.
 - C The median is 3.5.
 - D The median is 4.
 - E None of the above.
- 7 **MC** Find the mode of the following group of scores: 1, 1, 5, 2, 2, 1, 5, 4, 5
- A 1
 - B 5
 - C 3
 - D 2
 - E 1 and 5 are both modes.
- 8 Thirty men were asked to reveal the number of hours they spent doing housework each week. The results are detailed below.

1	5	2	12	2	6	2	8	14	18
0	1	1	8	20	25	3	0	1	2
7	10	12	1	5	1	18	0	2	2

- a Represent the data in a frequency distribution table. (Use a class size of 5.)
 - b Find the mean number of hours that the men spent doing housework.
 - c Find the median class for hours spent by the men at housework.
 - d Find the modal class for hours spent by the men at housework.
- 9  The data at right give the age of 25 patients admitted to the emergency ward of a hospital.
- | | | | | |
|----|----|----|----|----|
| 18 | 16 | 6 | 75 | 24 |
| 23 | 82 | 75 | 25 | 21 |
| 43 | 19 | 84 | 76 | 31 |
| 78 | 24 | 20 | 63 | 79 |
| 80 | 20 | 23 | 17 | 19 |
- a Represent the data in a frequency distribution table. (Use a class size of 15.)
 - b Draw a histogram of the data.
 - c What word could you use to describe the pattern of the data in this distribution?
 - d Find the mean age of patients admitted.
 - e Find the median class for age of patients admitted.
 - f Find the modal class for age of patients admitted.
 - g Draw an ogive of the data.
 - h Use the ogive to determine the median age. (The median is the middle score, or 50th percentile.)
 - i Do any of your statistics (mean, median or mode) give a clear representation of the typical age of an emergency ward patient?
 - j Give some reasons which could explain the pattern of the distribution of data in this question.
- 10 The batting scores for two cricket players over 6 innings are as follows:
- Player A 31, 34, 42, 28, 30, 41
- Player B 0, 0, 1, 0, 250, 0
- a Find the mean score for each player.
 - b Which player appears to be best, based upon mean result?
 - c Find the median score for each player.
 - d Which player appears to be best when the decision is based on the median result?
 - e Which player do you think would be the most useful to have in a cricket team and why? How can the mean result sometimes lead to a misleading conclusion?



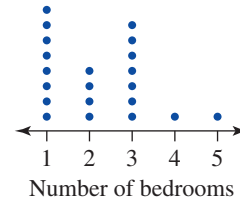
- 11 The resting pulse rate of 20 female athletes was measured. The results are detailed below.

50 52 48 52 71 61 30 45 42 48
43 47 51 52 34 61 44 54 38 40

- Construct a frequency distribution table.
- Find the mean of the data.
- Find the median class of the data.
- Find the modal class of the data.
- Draw an ogive of the data. (You may like to use a CAS calculator for this.)
- Use the ogive to determine the median pulse rate.

The following information relates to Questions 12 and 13.

The dot plot at right shows the distribution of the number of bedrooms in each of 21 apartments advertised for sale in a new high-rise apartment block.



- 12 **MC** The mode of this distribution is:
A 1 **B** 2 **C** 3 **D** 7 **E** 8
 [©VCAA 2007]
- 13 **MC** The median of this distribution is:
A 1 **B** 2 **C** 3 **D** 4 **E** 5
 [©VCAA 2007]
- 14 **MC** In a set of data there is one score which is extremely small when compared to all the others. This outlying value is most likely to:
A have greatest effect upon the mean of the data.
B have greatest effect upon the median of the data.
C have greatest effect upon the mode of the data.
D have a great effect on all of the statistics: mean, median and mode.
E have very little effect on any of the statistics as we are told that the number is extremely small.
- 15 The following frequency table gives the number of employees in different salary brackets for a small manufacturing plant.

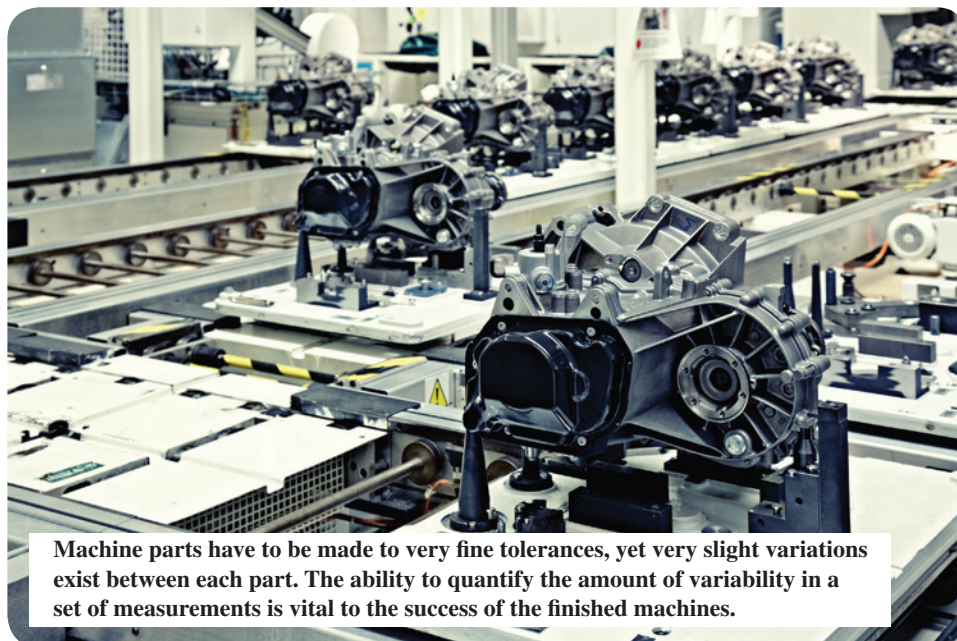
Position	Salary (\$)	Number of employees
Machine operator	24 000	50
Machine mechanic	26 000	15
Floor steward	30 000	10
Manager	75 000	4
Chief executive officer	120 000	1

- Workers are arguing for a pay rise but the management of the factory claims that workers are well paid because the mean salary of the factory is \$28 875. Are they being honest?
 - Suppose that you were representing the factory workers and had to write a short submission in support of the pay rise. How could you explain the management's claim? Quote some other statistics in favour of your case.
- 16 Design a set of five numbers with:
- mean = median = mode
 - mean > median > mode
 - mean > median = mode
 - median < mean = mode.

1E Measures of variability

So far we have been concerned with finding a piece of data which is representative or typical of a set of data — a piece that could be thought of as the centre of the data. Such numbers in themselves do not fully describe a set of data. It is also useful to be able to measure the amount or variability of spread in the data. Are the data all similar, or are they dispersed in their nature? In this section we will examine four methods by which the spread of data can be measured:

- the range
- the interquartile range
- the standard deviation
- the variance.



Machine parts have to be made to very fine tolerances, yet very slight variations exist between each part. The ability to quantify the amount of variability in a set of measurements is vital to the success of the finished machines.

The range

The *range* is the most basic of the measures of variability.

- It is found quite simply by taking the smallest score from the largest.

- **Range** = $X_{\max} - X_{\min}$

For example, for the set of data 3, 5, 4, 4, 2, 3, 0, 1, 4, 5 (which gave the number of pets in each of 10 households):

1. the smallest score = 0
2. the largest score = 5
3. the range = $5 - 0$
= 5

In the case of the grouped data it is necessary to assume that the greatest scored observation occurs at the upper end of the greatest class and the lowest score occurs at the lower end of the smallest class. Consider the following worked example.

WORKED EXAMPLE 10

The frequency distribution table below shows the heights of boys who are competing for places in a basketball squad. Find the range of the data.

Height (cm)	Frequency
170–<175	3
175–<180	6
180–<185	12

Height (cm)	Frequency
185–<190	10
190–<195	8
195–<200	1

THINK

The range is the difference between the highest and lowest scores. Take the highest score to be 200. The smallest score is the smallest possible score in the lowest class. That would make it 170.

WRITE

$$\begin{aligned}\text{Range} &= 200 - 170 \\ &= 30\end{aligned}$$

- The problem with using the range as a measure of variability is that it is severely affected by extreme values.

Take the pets data as an example. We might have surveyed 1000 houses and found that all gave similar data to the ten scores listed — that is, a pets number of between 0 and 5, but then we might just strike

that one house which harboured 99 cats. The range becomes $99 - 0 = 99$. This would indicate that there is a large degree of variability in the number of pets that people kept, which would be a misleading statistic because in reality the vast majority of the houses showed only a small variation (range = 5).

The interquartile range

One way to overcome the problems caused by extreme values is to exclude the top and bottom quarters of the data and then consider the range formed by the remaining data. It is necessary to find the lower quartile (denoted Q_1) and the upper quartile (denoted Q_3).

- The *lower quartile* (Q_1) is the piece of data that is $\frac{1}{4}$ of the way through the distribution (when it is arranged in order of size). The lower quartile is the 25th percentile of the data.
- The *upper quartile* (Q_3) is the piece of data that is $\frac{3}{4}$ of the way through the distribution. It is the 75th percentile of the data.
- **The interquartile range, IQR, is the difference between the upper and lower quartiles:**

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

The upper and lower quartiles and the interquartile range can be found by using the following steps:

- Step 1.** Arrange the data in order of size.
- Step 2.** Divide the data into halves by finding the median. If there is an odd number of scores then the median will be one of the original scores. In this case it should not be included in either the lower half or the upper half of the scores. If there is an even number of scores the median will lie halfway between two scores and will divide the data neatly into two equal sets.
- Step 3.** Find the lower quartile by locating the median of the lower half of the data.
- Step 4.** Find the upper quartile by locating the median of the upper half of the data.
- Step 5.** Find the interquartile range by calculating the difference between the upper and lower quartiles.

WORKED EXAMPLE 11

Determine the interquartile range of the following data:

12, 9, 4, 6, 5, 8, 9, 4, 10, 2.

THINK

- 1 First arrange the data in order of size.
- 2 Find the position of the median. The median is the $\frac{n+1}{2}$ th score.

2, 4, 4, 5, 6, 8, 9, 9, 10, 12

↑
median

- 3 The data have been divided into halves, each containing 5 values. The lower quartile is the median of the lower half. The upper quartile is the median of the upper half.

- 4 The interquartile range is the difference between the upper and lower quartiles.

WRITE

2, 4, 4, 5, 6, 8, 9, 9, 10, 12

The median is the $\frac{10+1}{2}$ th score — that is, the 5.5th score.

That is, halfway between the 5th and 6th scores.

The position of upper and lower half

medians is the $\frac{5+1}{2}$ th score — that is, the 3rd score.

2, 4, 4, 5, 6, | 8, 9, 9, 10, 12

$Q_1 = 4$

$Q_3 = 9$

$$\begin{aligned}\text{IQR} &= Q_3 - Q_1 \\ &= 9 - 4 \\ &= 5\end{aligned}$$

It is quite difficult to find the interquartile range of grouped data. The easiest approach is to draw an ogive of the data and use it to find the 75th and 25th percentiles.

WORKED EXAMPLE 12

eBookplus

The following frequency distribution table gives the number of customers who order different volumes of concrete from a readymix concrete company during the course of a day. Find the interquartile range of the data.

TUTORIAL
eles-1340
Worked example 12

Volume (m ³)	Frequency
0.0–	15
0.5–	12
1.0–	10

Volume (m ³)	Frequency
1.5–	8
2.0–	2
2.5–	4

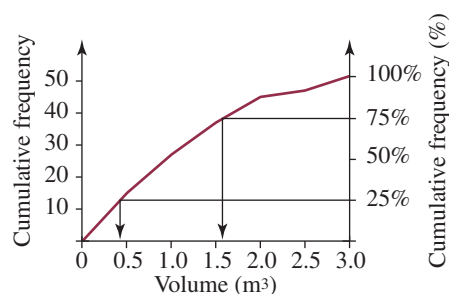
THINK

- To find the 25th and 75th percentiles from the ogive, first add a cumulative frequency column to the frequency distribution table and fill it in.

WRITE/DRAW

Volume	<i>f</i>	cf
0.0–	15	15
0.5–	12	27
1.0–	10	37
1.5–	8	45
2.0–	2	47
2.5–	4	51

- Draw the ogive. A percentage axis will be useful.



- Find the upper quartile (75th percentile) and lower quartile (25th percentile) from the ogive.

$$Q_3 = 1.6 \text{ m}^3$$

$$Q_1 = 0.4 \text{ m}^3$$

- The interquartile range is the difference between the upper and lower quartiles.

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 1.6 - 0.4 \\ &= 1.2 \text{ m}^3 \end{aligned}$$

The standard deviation

The *standard deviation* is the most sophisticated and the most useful of the measures of spread.

The standard deviation, *s*, can be calculated by using the following formula:

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

To calculate the standard deviation by hand, use the following steps:

- Step 1. Find the mean.
- Step 2. Find the difference between each piece of data and the mean.
- Step 3. Square the differences.

Step 4. Add the squared differences.

Step 5. Divide by the number of scores less one.

Step 6. Take the square root.

This algorithm is used to find the standard deviation in the following worked example.

WORKED EXAMPLE 13

The following data give the number of lollies in each of 8 packets. Find the standard deviation of the data.

14, 14, 13, 15, 16, 13, 14, 17

THINK

- 1 Find the mean.
- 2 Find the difference between each score and the mean.
- 3 Square each difference.
- 4 Add the squared differences.

WRITE

$$\bar{x} = \frac{14+14+13+15+16+13+14+17}{8}$$

$$= 14.5$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
14	-0.5	0.25
14	-0.5	0.25
13	-1.5	2.25
15	0.5	0.25
16	1.5	2.25
13	-1.5	2.25
14	-0.5	0.25
17	2.5	6.25
Total		14

- 5 Divide by the number of scores less 1. (Number of scores is 8; $8 - 1 = 7$; so divide by 7.)

$$14 \div 7 = 2$$

- 6 Take the square root and round to 4 decimal places.

$$\sqrt{2} = 1.4142$$

The standard deviation $s = 1.4142$.

Fortunately, you will not always have to go through this series of steps each time you wish to calculate a standard deviation. Your calculator should have a built-in program for the computation of standard deviations.

A final point worth noting about the standard deviation is that, despite its sophistication, it is still influenced to a high degree by extreme values. Care should be taken when using this statistic with data that include such values.

The variance

The final commonly used measure of spread is the variance.

The variance is the standard deviation squared.

- The symbol s^2 is often used for variance and the steps used in its calculation can be summarised by the formula:

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

- To find the variance without the use of the calculator's program, use the same steps listed as an algorithm for calculating the standard deviation, but stop at step 5.
- If the variance is to be found using the calculator's program, first find the standard deviation, then square it.

**WORKED EXAMPLE 14**

The following frequency distribution gives the prices paid by a car wrecking yard for 40 car wrecks.

eBookplus

TUTORIAL
eles-1341
Worked example 14

Price (\$)	Frequency
0—<500	2
500—<1000	4
1000—<1500	8
1500—<2000	10

Price (\$)	Frequency
2000—<2500	7
2500—<3000	6
3000—<3500	3

Use a calculator to:

- a find the mean and standard deviation in the price paid for these wrecks
- b find the variance in prices of car wrecks.

THINK

- a 1 Enter the midpoint of each price range, and the corresponding frequency into your calculator.
- 2 Determine the standard deviation of the data.

b Variance = (standard deviation)² = s²

WRITE

a Standard deviation = \$797.03

b Variance = (797.03)²
= 635 256.82

Exercise 1E Measures of variability

- 1 Find the range of each of the following sets of data.

- a 2, 5, 4, 5, 7, 4, 3
- b 103, 108, 111, 102, 111, 107, 110
- c 2.5, 2.8, 3.4, 2.7, 2.6, 2.4, 2.9, 2.6, 2.5, 2.8
- d 3.20, 3.90, 4.25, 7.29, 1.45, 2.77, 8.39

- 2 **MC** The percentage profits of the new business for the first 9 months of its operation are as follows:

−5.4%, −4.2%, 3.6%, 2.9%, 0.7%, −5.2%, −3.9%, 0.8%, −1.1%

The range of the profits is:

- A −9.0%
- B 2.8%
- C 2.9%
- D 8.3%
- E 9.0%

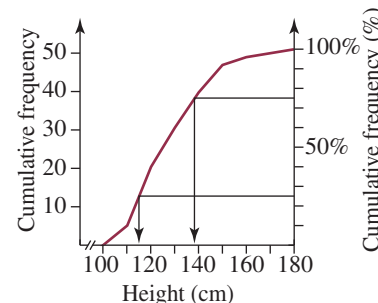
- 3 **WE10** The frequency distribution table below shows the weights of women joining a weight-watchers program. Find the range of the data.

Weight	60—<65	65—<70	70—<75	75—<80	80—<85	85—<90	90—<95
Frequency	3	5	10	13	12	8	4

- 4 **WE11** Find the interquartile range of the following data.

- a 3, 4, 4, 5, 7, 9, 10
- b 17, 20, 19, 25, 29, 27, 28, 25, 29
- c 52, 55, 53, 53, 54, 55, 52, 53, 54, 52
- d 12, 14, 15, 12, 14, 19, 17, 15, 18, 20

- 5 Use the ogive at right to determine the interquartile range of the data.



- 6 **WE12** The following frequency distribution table gives the amount of time spent by 50 people on shopping for Christmas presents. Find the IQR of the data.

Time (h)	0–	0.5–	1–	1.5–	2–	2.5–	3–	3.5–
Frequency	1	2	7	15	13	8	2	2

- 7 **MC** Calculate the interquartile range of the following data:

17, 18, 18, 19, 20, 21, 21, 23, 25

- A 8 B 18 C 4 D 20 E None of these
- 8 **WE13** Use the algorithm (series of steps) to find the standard deviation of the following data.
- a 3, 5, 8, 2, 7, 1, 6, 5 b 11, 8, 7, 12, 10, 11, 14
- c 25, 15, 78, 35, 56, 41, 17, 24 d 5.2, 4.7, 5.1, 12.6, 4.8

- 9 Use a calculator to check each of the standard deviations that you calculated in the previous question.

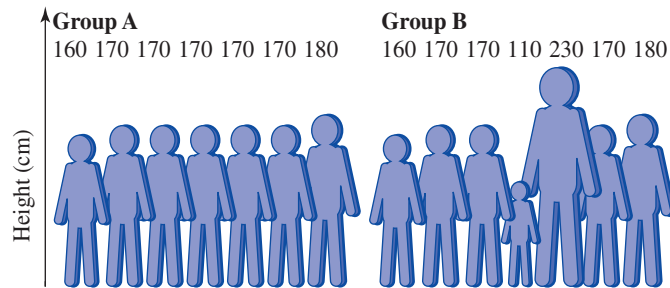
The following information relates to Questions 10 and 11.

The mean weight for the team of eight people is 69 kg; the standard deviation of the weights of these people is 3.4 kg.

- 10 **MC** What is the combined weight of the team?
- A 72.4 kg B 524.8 kg C 530.4 kg D 552 kg E 580 kg
- 11 **MC** The team is about to play a laser skirmish. Before the game begins, everyone needs to put on a special vest that weighs 1.8 kg. The effective weight of each person in the team is now the sum of their own weight and the weight of the vest.
- The effective weights of the team have:
- A a mean of 69 kg and a standard deviation of 3.4 kg
- B a mean of 69 kg and a standard deviation of 5.2 kg
- C a mean of 70.8 kg and a standard deviation of 3.4 kg
- D a mean of 70.8 kg and a standard deviation of 5.2 kg
- E a mean of 70.8 kg and a standard deviation of 1.6 kg

- 12 Consider the following two groups of people.

- a Calculate the mean height, median height and mode height for each group. What do you notice?
- b Are the groups really the same?
- c Which group would you expect to show the greatest range in heights?
- d Which group would you expect to show the greatest interquartile range in heights?
- e Which group would you expect to show the greatest standard deviation in heights?
- f Calculate these statistics to confirm your predictions.



- 13 **MC** The interquartile range is considered to be a better measure of the variability of a set of scores than the range because:
- A it takes into account more scores
- B it is the difference between the upper and lower quartiles
- C it is easier to calculate
- D it is centred around the median
- E it is not affected by extreme values
- 14 The following frequency distribution table shows the number of visitors that came to a city museum during the course of a month.

Visitor number	80–	90–	100–	110–	120–	130–
Frequency	1	4	11	9	4	2

- a Find the range of the data.
- b Find the mean of the data.
- c Find the standard deviation and variance of the data.

- 15 **MC** Calculate the standard deviation of the following data to 3 decimal places.

Score	10–	20–	30–	40–	50–
Frequency	1	6	9	4	1

- A 3.027 B 9.209 C 9.437 D 34.048 E None of these

- 16 **WE14** The following frequency distribution table shows the life expectancy of 175 household light globes.

Life (h)	Frequency
200–<250	2
250–<300	5
300–<350	12
350–<400	25
400–<450	42

Life (h)	Frequency
450–<500	38
500–<550	26
550–<600	15
600–<650	7
650–<700	3

- a Use a calculator to find the mean and standard deviation in the lifetimes of these light globes.
b Find the variance in the lifetimes of these light globes.

- 17 **MC** Which of the following gives a correct method for calculating the variance of a set of data?

- A Average the difference that each score is from the mean.
B Square the standard deviation.
C Take the square root of the standard deviation.
D Average the upper and lower quartiles.
E None of the above

- 18 The following frequency distribution table shows the distribution of daily maximum temperatures during the course of a full year.

Maximum temperature (°C)	Number of days
0–	4
5–	22
10–	95
15–	124

Maximum temperature (°C)	Number of days
20–	94
25–	19
30–	5
35–	2

- a Add a cumulative frequency column to the table.
b Draw an ogive of the data.
c Find the upper and lower quartiles of the data and calculate the interquartile range.
d Use the ogive to find the median (50th percentile of the data).
e Find the mean of the data.
f Find the standard deviation of the data.
g Find the variance of the data.
h Find the range of the data.

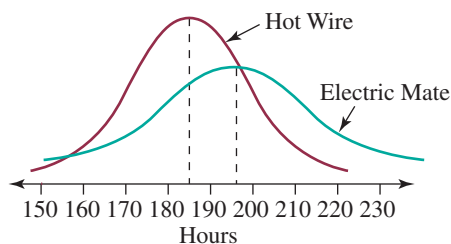
- 19 The following data give the number of fruit that have formed on each of 30 trees in an orchard.

45	48	52	36	38	72	36	74	56	46
81	73	46	48	44	39	52	58	57	65
60	53	54	58	41	44	47	76	68	55

- a Complete a frequency distribution table for the data with the class intervals being 30–39, 40–49, and so on.
b Draw an ogive of the data.
c Use the ogive to find the median, lower quartile and upper quartile of the data.
d Find the interquartile range of the data.
e Find the mean of the data.
f Find the standard deviation of the data.
g Find the range of the data.

- 20 The curves drawn at right show the lifetimes of two samples of different brands of toaster elements when subject to continued use.

- Which brand has the longest mean life?
- Estimate the mean life of each brand.
- Which brand has the greatest standard deviation in its performance?
- What does this say about the consistency of this element?
- Which brand is best? Give a case in support of Electric Mate. Give a case in support of Hot Wire.



- 21 Crunch and Crinkle are two brands of potato crisps. Each is sold in packets nominally of the same size and for the same price. Upon investigation of a sample of packets of each it is found that Crunch and Crinkle have the same mean weight (25 g). The standard deviation of the weights of Crunch is, however, 5 g and the standard deviation of the weights of Crinkle is 2 g. Which brand do you think would represent the best value for money under these circumstances? Why?



eBookplus

DIGITAL DOC
doc-9531
WorkSHEET 1.2

1F Stem-and-leaf plots

As an alternative to a frequency distribution table a *stem-and-leaf plot* (also called a *stemplot*) may be used to group and summarise data.

WORKED EXAMPLE 15

The data at right show the weights (in kg) of 20 possums trapped, weighed, then released by a wildlife researcher.

Prepare a stem-and-leaf diagram for the data.

1.8	0.9	0.7	1.4	1.6
2.1	2.7	2.2	1.8	2.3
2.3	1.5	1.1	2.2	3.0
2.5	2.7	3.2	1.9	1.7

THINK

- Each data element is made up of two parts: a stem and a leaf. The first weight was 1.8 kg. The stem of this number could be considered to be 1 and the leaf 8. The second value was 0.9. It has a stem of 0 and a leaf of 9. To compose the stem-and-leaf plot, rule a vertical column of stems, then enter the leaf of each value in a neat row beside the appropriate stem. The first row of the stem-and-leaf plot is recording all data from 0.0 to 0.9. The second row is recording data from 1.0 to 1.9, and so on. Attach a key to the plot to show the reader the meaning of each entry.

WRITE

Key: 0|7 = 0.7 kg

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

- It is a convention to assemble the data in order of size, so this stem-and-leaf plot should be rewritten in such a way that the numbers in each row of 'leaves' are in ascending order.

Key: 0|7 = 0.7 kg

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

- In practice it may be best when forming the plot to record the leaves initially in pencil so that they can be easily erased and the plot presented with leaves in proper order.
- Note that every stem-and-leaf plot must have a key attached to it that leaves the reader in no doubt as to the meaning of each entry.

The key in the worked example 14 ($0|7 = 0.7 \text{ kg}$) could be read as ‘stem 0 and leaf 7 means 0.7 kg’. The key is important because there are several other ways that this information could have been presented. Figure 1 below is an example of another equally correct way of presenting the ‘possum’ data on a stem-and-leaf diagram.

- Although the presentation of the data is different it is quite an acceptable plot as the reader is left in no doubt as to the meaning of each entry.
- When preparing a stem-and-leaf plot, try to keep the numbers in neat vertical columns because a neat plot gives the reader an idea of the distribution of scores. The plot itself looks a bit like a histogram turned on its side.
- It is also useful to be able to represent data with a class size of 5. This could be done for the same data by choosing stems 0^* , 1 , 1^* , 2 , 2^* , 3 (see figure 2 below). Here the class with stem 1 contains all the data from 1.0 to 1.4 and the stem 1^* contains the data from 1.5 to 1.9 and so on. If stems are split in this way it is a good idea to include two entries in the key.

Key: $0.0|7 = 0.7 \text{ kg}$

Stem	Leaf
0.0	7 9
1.0	1 4 5 6 7 8 8 9
2.0	1 2 2 3 3 5 7 7
3.0	0 2

Figure 1

Key: $1|1 = 1.1 \text{ kg}$
 $1^*|5 = 1.5 \text{ kg}$

Stem	Leaf
0^*	7 9
1	1 4
1^*	5 6 7 8 8 9
2	1 2 2 3 3
2^*	5 7 7
3	0 2

Figure 2

A stem-and-leaf plot has the following advantages over a frequency distribution table:

1. The plot itself gives a graphical representation of the spread of data. (It is rather like a histogram turned on its side.)
2. All the original data are retained so there is no loss of accuracy when calculating statistics like the mean and standard deviation. With a frequency distribution table, some generalisations are made when these values are calculated.

WORKED EXAMPLE 16

The following data give the length of gestation (in days) for 24 mothers. Prepare a stem-and-leaf plot of the data using a class size of 5.

280	287	285	276	266	292
288	273	295	279	284	271
292	288	279	281	270	278
281	292	268	282	275	281

THINK

- 1 A group size of 5 is required. The smallest value is 266 and the largest is 295, so make the stems:

26^* , 27, 27^* , 28, 28^* , 29, 29^* .

The key should give a clear indication of the meaning of each entry.

- 2 Enter the data piece by piece. Enter the leaves in pencil at first so that they can be rearranged into order of size. Check that there have been 24 pieces of data entered.

WRITE

Key: $26^*|6 = 266 \text{ days}$
 $27|0 = 270 \text{ days}$

Stem	Leaf
26^*	6 8
27	3 1 0
27^*	6 9 9 8 5
28	0 4 1 1 2 1
28^*	7 5 8 8
29	2 2 2
29^*	5

- 3 Now rearrange the leaves into order of size.

Key: $26^*|6 = 266$ days

$27|0 = 270$ days

Stem	Leaf
26 *	6 8
27	0 1 3
27 *	5 6 8 9 9
28	0 1 1 1 2 4
28 *	5 7 8 8
29	2 2 2
29 *	5

WORKED EXAMPLE 17

Find the interquartile range of the data presented in the following stem-and-leaf plot.

Key: $15|4 = 154$

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

THINK

- 1 Count the number of pieces of data and work out the position of the median.

WRITE

There are 30 pieces of data so the median

will be the $\frac{30+1}{2}$ th score — that is, the

15.5th score, or halfway between the 15th and 16th scores.

- 2 Now find the position of the lower and upper quartiles. The lower quartile is the median of the lower half of the data. The upper quartile is the median of the upper half. Each half consists of 15 scores.

Q_1 will be the $\frac{15+1}{2}$ th score in the lower

half — that is, the 8th score in the lower half.

Q_3 will be the 8th score in the upper half of the data.

- 3 Count through the data to find the position of the medians of each half and mark these positions on the stemplot.

Key: $15|4 = 154$

Stem	Leaf	Q_1	Median
15	4 8 8		
16	1 3 3 6	⑧	
17	0 0 1 4 7 9 9		① 9
18	1 2 3 3 5 7		⑧ 8 9
19	2 7 8		
20	0 2		

- 4 Write the values of upper and lower quartiles.

$$Q_1 = 168$$

$$Q_3 = 188$$

- 5 The interquartile range is the difference between the upper and lower quartiles.

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 188 - 168 \\ &= 20 \end{aligned}$$

Exercise 1F Stem-and-leaf plots

- 1 **WE15** The data below give the number of errors made each week by 20 machine operators. Prepare a stem-and-leaf diagram of the data using stems of 0, 10, 20, etc.

6	15	20	25	28
18	32	43	52	27
17	26	38	31	26
29	32	46	13	20

- 2 The data below give the time taken for each of 40 runners on a 10-km fun run. Prepare a stem-and-leaf diagram for the data using a class size of 10.

36	42	52	38	47	59	72	68	57	82
66	75	45	42	55	38	42	46	58	39
42	58	40	41	47	53	68	43	39	48
71	42	50	46	40	52	37	54	48	52

- 3 **WE16** The typing speed of 30 word processors is recorded below. Prepare a stem-and-leaf diagram of the data using a class size of 5.

96	102	92	96	95	102	95	115	110	108
88	86	107	111	107	108	103	121	107	96
124	95	98	102	108	112	120	99	121	130

- 4 Twenty transistors are tested with increasing voltage until they are destroyed. The maximum voltage that each could withstand is recorded below. Prepare a stem-and-leaf plot of the data using a class size of 0.5.

14.8	15.2	13.8	14.0	14.8
15.7	15.5	15.6	14.7	14.3
14.6	15.2	15.9	15.1	14.3
14.6	13.9	14.7	14.5	14.2

- 5 **WE17**  The following stem-and-leaf plot gives the exact weight of 24 packets of biscuits. Find the interquartile range of the data.

Key: 248|4 = 248.4 g

Stem	Leaf
248	4 7 8
249	2 3 6 6
250	0 0 1 1 6 9 9
251	1 5 5 5 6 7
252	1 5 8
253	0

- 6  The time taken for a test vehicle to accelerate from 0 to 100 km/h is recorded during a test of 24 trials. The results are represented by the stem-and-leaf diagram at right.

- a Find the median of the data.
b Find the upper and lower quartiles of the data.
c Find the interquartile range of the data.

Key: 7|2 = 7.2 seconds
7*|6 = 7.6 seconds

Stem	Leaf
7	2 4 4
7*	5 5 7 9
8	0 0 1 2 4 4 4
8*	5 5 6 8 9
9	2 2 3
9*	0 2

Questions 7 to 10 refer to the stem-and-leaf plot below.

Key: 12|1 = 1210
12*|5 = 1250

Stem	Leaf
12	1 2 4
12*	5 7 7 9 9
13	0 1 1 2 3 4 4
13*	5 6 6 7 9 9
14	0 2 3 4
14*	0 1

- 7 **MC** The class size used in the stem-and-leaf plot is:
A 50 **B** 5 **C** 10 **D** 1 **E** none of these
- 8 **MC** The number of scores that have been recorded:
A is 27 **B** is 33 **C** is 1410 **D** is 1210
E cannot be determined without seeing the original data
- 9 **MC** The median of the data is:
A 14 **B** 1340 **C** 13.4 **D** 1335 **E** none of these
- 10 **MC** The interquartile range of the data is:
A 14 **B** 1390 **C** 1290 **D** 100 **E** none of these
- 11 The maximum hand span (in cm) of 20 male concert pianists is recorded as follows:

23.6	20.2	22.8	21.4	25.1
24.8	23.2	21.6	20.7	23.6
22.8	24.6	21.8	22.8	23.1
24.6	21.7	24.7	22.2	23.0

- a** Complete a stem-and-leaf plot to represent the data.
b Find the median of the data.
c Find the upper and lower quartile of the data.
d Find the interquartile range of the data.

- 12 The heights (in cm) of a sample of 30 plants are recorded as follows.

93	88	94	99	91
98	96	117	101	97
114	84	96	103	108
107	85	126	110	111
92	101	132	103	82
115	90	110	126	85

- a** Complete a stem-and-leaf plot to represent the data.
b Find the median of the data.
c Find the upper and lower quartile of the data.
d Find the interquartile range of the data.



1G Boxplots

Five-number summary

- A **five-number summary** is a list consisting of the lowest score ($X_{\min}$), lower quartile (Q_1), median (Q_2), upper quartile (Q_3) and greatest score ($X_{\max}$) of a set of data.
- A five-number summary gives information about the spread of a set of data.
- The convention is not to detail the numbers with labels but to present them in order; so, for example, the five-number summary:

4 15 21 23 28

would be interpreted as lowest score 4, lower quartile 15, median 21, upper quartile 23 and greatest score 28.

WORKED EXAMPLE 18

From the following five-number summary find:

- a** the median **b** the interquartile range **c** the range.

29 37 39 44 48

THINK

The figures are presented in the order of lowest score, lower quartile, median, upper quartile, greatest score.

WRITE

$X_{\min} = 29$, $Q_1 = 37$, median = 39,
 $Q_3 = 44$, $X_{\max} = 48$

eBookplus

INTERACTIVITY
 int-0802
 Boxplots
 Five-number summary



a The median is 39.

b The interquartile range is the difference between the upper and lower quartiles.

c The range is the difference between the greatest score and the lowest score.

a Median = 39

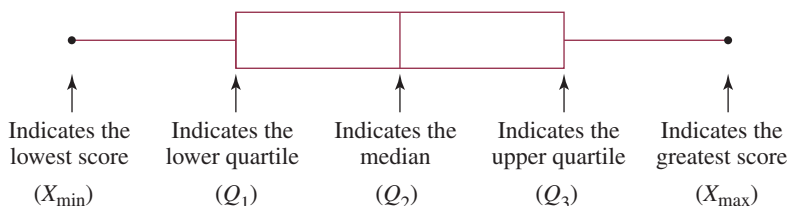
$$\begin{aligned} \text{b IQR} &= Q_3 - Q_1 \\ &= 44 - 37 \\ &= 7 \end{aligned}$$

$$\begin{aligned} \text{c Range} &= X_{\max} - X_{\min} \\ &= 48 - 29 \\ &= 19 \end{aligned}$$

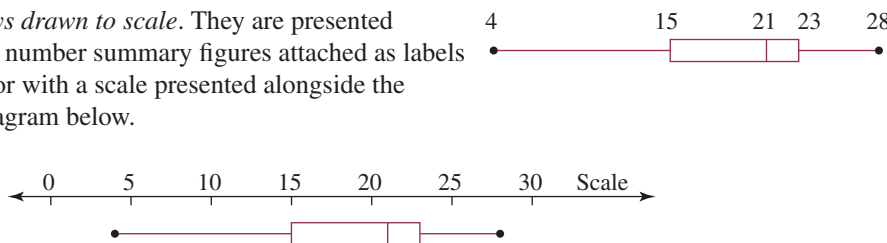
Boxplots

A **boxplot** (or **box-and-whisker plot**) is a graph of the five-number summary.

- It is a powerful way to show the spread of data.
- Boxplots consist of a central divided box with attached 'whiskers'.
- The box spans the interquartile range.
- The median is marked by a vertical line inside the box.
- The whiskers indicate the range of scores:



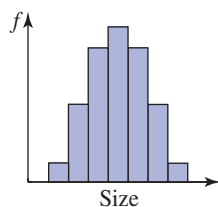
- Boxplots are *always drawn to scale*. They are presented either with the five number summary figures attached as labels (diagram at right) or with a scale presented alongside the boxplot like the diagram below.



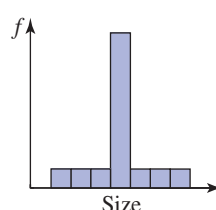
Interpreting a boxplot

- The boxplot neatly divides the data into four sections.
- One-quarter of the scores lie between the lowest score and the lower quartile, one-quarter between the lower quartile and the median, one-quarter between the median and the upper quartile, and one-quarter between the upper quartile and the greatest score.
- The reader can easily see where the data cluster occurs. For example, a small box with relatively long whiskers would indicate that half of the data (from Q_1 to Q_3) would be confined to a small range and the data could be described as 'clustered'. A wide box with relatively short whiskers would indicate that half of the data (from Q_1 to Q_3) would be spread over a wide range and the data could be described as 'spread'.

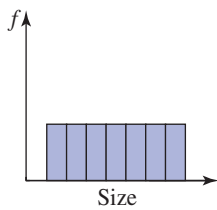
Consider the following boxplots with their matching histograms.



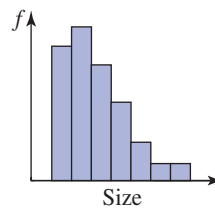
Normally distributed data



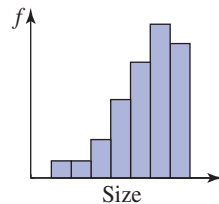
Clustered data



Spread data



Positively skewed data



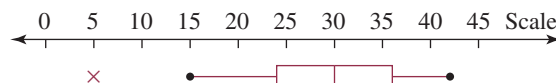
Negatively skewed data



Identification of extreme values

- Extreme values often make the whiskers appear longer than they should and hence give the appearance that the data are spread over a much greater range than they really are.
- If an extreme value occurs in a set of data it can be denoted by a small cross on the boxplot. The whisker is then shortened to the next largest (or smallest) figure.

The boxplot below shows that the lowest score was 5. This was an extreme value as the rest of the scores were located within the range 15 to 42.



- Extreme values can be determined by using a formula which will be covered in *Further Mathematics* Units 3 and 4. For now, extreme values are generally estimated by close examination of the data or of the graph.

WORKED EXAMPLE 19



The following stem-and-leaf plot gives the speed of 25 cars caught by a roadside speed camera.

- Prepare a five-number summary of the data.
- Draw a boxplot of the data. (Identify any extreme values.)
- Describe the distribution of the data.

Key: 8|2 = 82 km/h

8*|6 = 86 km/h

Stem	Leaf
8	2 2 4 4 4 4
8*	5 5 6 6 7 9 9 9
9	0 1 1 2 4
9*	5 6 9
10	0 2
10*	
11	4

THINK

- First determine the median of the data. The scores are in order, so the median is the $\frac{25+1}{2}$ th score, i.e., the 13th score. Count 13 scores from the lowest score.
- Determine the value of the lower quartile, which is in the middle of the lower half of the scores.

WRITE/DRAW

- The median of the data is the 13th score, which is 89.

There are 12 scores in the lower half of the data, so the lower quartile is the $\frac{12+1}{2}$ th score

$$Q_1 = \frac{84 + 85}{2} = 84.5$$





- 3 The upper quartile is 6.5 scores from the highest value.

The upper quartile is 6.5 scores from the highest value.

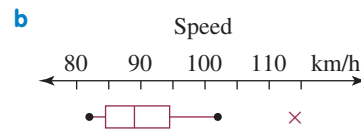
$$Q_3 = \frac{94 + 95}{2} = 94.5$$

- 4 Write the five-number summary.

Five-number summary:

$X_{\min}$	Q_1	Med	Q_3	$X_{\max}$
82	84.5	89	94.5	114

- b Start by ruling a suitable scale. Remember to include the units of measurement. The box represents the interquartile range so it runs from 84.5 to 94.5. The median is a vertical line in the box at 89. The whiskers should extend to the lowest score (82) and the highest score (114). But the score 114 is a great deal higher than any of the others in the set and might be regarded as an extreme value. It should be indicated by a cross and the whisker will extend only as far as 102 (the second largest number in the set).



- c Even when the extreme value is excluded the data appear to be skewed with high values being spread over a much greater range.
- c The data are skewed (positively) and include one extremely high value.

Exercise 1G Boxplots

- 1 **WE18** From the following five-number summary find:

a the median

b the interquartile range
6, 11, 13, 16, 32

c the range.

- 2 From the following five-number summary find:

a the median

b the interquartile range
101, 119, 122, 125, 128

c the range.

- 3 From the following five-number summary find:

a the median

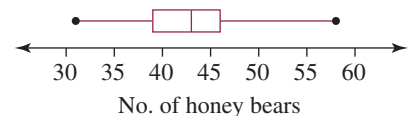
b the interquartile range
39.2, 46.5, 49.0, 52.3, 57.8

c the range.

- 4 The boxplot below shows the distribution of final points scored by a football team over a season's roster.

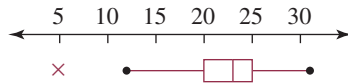


- a What was the team's greatest points score?
- b What was the team's least points score?
- c What was the team's median points score?
- d What was the range of points scored?
- e What was the interquartile range of points scored?
- 5 The boxplot at right shows the distribution of data formed by counting the number of honey bears in each of a large sample of packs.
- a What was the largest number of honey bears in any pack?
- b What was the smallest number of honey bears in any pack?
- c What was the median number of honey bears in any pack?



- d What was the range of numbers of honey bears per pack?
- e What was the interquartile range of honey bears per pack?

Questions 6 to 8 refer to the following boxplot.



- 6 **MC** The median of the data is:
 A 5 B 20 C 23 D 25 E 31
- 7 **MC** The interquartile range of the data:
 A is 23 B is 26 C is 5 D is 20 to 25
 E cannot be determined because of extreme values
- 8 **MC** Which of the following is not true of the data represented by the boxplot?
 A One-quarter of the scores are between 5 and 20.
 B Half of the scores are between 20 and 25.
 C The lowest quarter of the data is spread over a wide range.
 D Most of the data are contained between the scores of 5 and 20.
 E The data are skewed left.

eBookplus
DIGITAL DOC
 doc-9532
 Boxplots

- 9 The number of sales made each day by a salesperson is recorded over a 2-week period:

25, 31, 28, 43, 37, 43, 22, 45, 48, 33

- a Prepare a five-number summary of the data. (There is no need to draw a stem-and-leaf plot of the data. Just arrange it in order of size.)
- b Draw a boxplot of the data.

- 10 The data below show monthly rainfall in millimetres.

J	F	M	A	M	J	J	A	S	O	N	D
10	12	21	23	39	22	15	11	22	37	45	30

- a Prepare a five-number summary of the data.
- b Draw a boxplot of the data.

- 11 **WE19** The stemplot below details the age of 25 offenders who were caught during random breath testing.

Key: 1|8 = 18 years

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

- a Prepare a five-number summary of the data.
- b Draw a boxplot of the data.
- c Describe the distribution of the data.

- 12 The following stem-and-leaf plot details the price at which 30 houses in a particular suburb were sold.

Key: 12|4 = \$124 000

Stem	Leaf
12	4 7 9
13	0 0 2 5 5
14	0 0 2 3 5 5 7 9 9
15	0 0 2 3 7 7 8
16	0 2 2 5 8
17	5

- a Prepare a five-number summary of the data.
- b Draw a boxplot of the data.

- 13** The following data detail the number of hamburgers sold by a fast food outlet every day over a 4-week period.

M	T	W	T	F	S	S
125	144	132	148	187	172	181
134	157	152	126	155	183	188
131	121	165	129	143	182	181
152	163	150	148	152	179	181

- a** Prepare a stem-and-leaf plot of the data. (Use a class size of 10.)
b Draw a boxplot of the data.

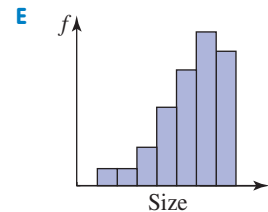
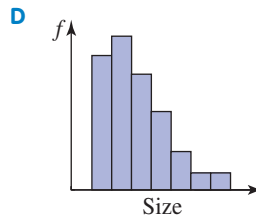
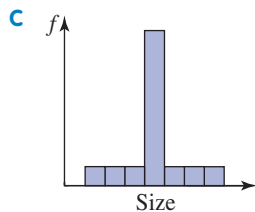
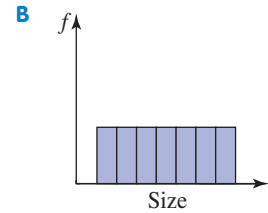
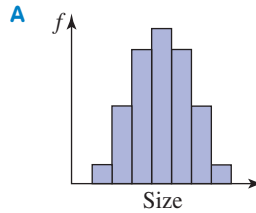


- 14** The following data show the ages of 30 mothers upon the birth of their first baby.

22	21	18	33	17	23	22	24	24	20
25	29	32	18	19	22	23	24	28	20
31	22	19	17	23	48	25	18	23	20

- a** Prepare a stem-and-leaf plot of the data. (Use a class size of 5.)
b Draw a boxplot of the data. Indicate any extreme values appropriately.
c Describe the distribution in words. What does the distribution say about the age at which mothers have their first baby?

- 15 MC** Match the boxplot with its most likely histogram.



1H Comparing sets of data

Back-to-back stem-and-leaf plots

Some of the most useful and interesting statistical investigations involve the comparison of two sets of data.

- *Back-to-back stem-and-leaf plots* are useful for comparing the distribution of two similar sets of data. This is particularly useful in the situation of controlled experiments.
- The two sets of data use the same central stem.
- One set of leaves is set to the right of the stem and the other to the left.
- Care must be taken when arranging the data of the left set. Place the smallest numeral closest to the central margin, then range outwards as the data size increases.
- The key generally relates to data that are presented on the right of the plot.

An example of such a plot is presented below. The data show the lifetimes of a sample of 40 batteries of each of two brands when fitted into a standard children's toy. Some of the toys are fitted with an ordinary brand battery and some with Brand X. Which battery is better?



Key: 6|9 = 69 hours

Ordinary brand Leaf	Stem	Brand X Leaf
8 6 2 0 0	6	9
9 9 9 8 8 6 4 0	7	3 5
8 8 7 5 3 1 1 1 0	8	2 4 8
9 6 6 4 2 2 2 0 0	9	0 1 4 5 5 9
8 7 5 3 1 1 1	10	0 0 2 5 8 8 9 9
4 2	11	0 0 1 1 3 3 6 7 9
	12	1 4 6 6 6 7 8 8
	13	3 5
	14	6

The spread of each set of data can be seen graphically from the stem-and-leaf plot. In this case it can be seen that, although Brand X showed a little more variability than the ordinary brand, it generally gave a longer lasting performance.

Parallel boxplots

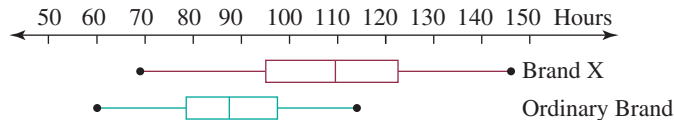
- Two or more sets of data may be compared by using *parallel* or *side-by-side boxplots*.
- The boxplots share a common scale.
- Numerical comparisons can be made between the sets of data based upon the size and position of the range, interquartile range and median. This is a strong feature of a boxplot.
- In general, a histogram or stem-and-leaf plot is better than a boxplot at giving the reader information about the distribution of a set of scores, but boxplots have greater scope for making quantitative comparisons.

In the case of the battery test data above, the following side-by-side boxplot would result. (Quartiles and medians are found in the usual way.)

We can make the following

comparisons between the sets of data:

- Brand X showed more variability in its performance (that is, its lifetime) than the ordinary brand. (Brand X range = 77, ordinary brand range = 54; Brand X interquartile range = 27.5, ordinary brand interquartile range = 19).
- The longest lifetime recorded was that of a Brand X battery (146).
- The shortest lifetime recorded was that of an ordinary brand battery (60).
- Brand X batteries' median lifetime was better than that of the ordinary brand (Brand X median = 109.5, ordinary brand median = 87.5).
- Over one-quarter of Brand X batteries were better performers than the best ordinary brand battery (that is, had longer lifetimes than the longest of the ordinary brand batteries' lifetimes).



WORKED EXAMPLE 20

The stem-and-leaf plot below shows the weights of two samples of chickens 3 months after hatching. One group of chickens (Group A) had been given a special growth hormone. The other group (Group B) was kept under identical conditions but was not given the hormone. Prepare side-by-side boxplots of the data and draw conclusions about the effectiveness of the growth hormone.

Key: 0*|8 = 0.8 kg
1|3 = 1.3 kg

Group B Leaf	Stem	Group A Leaf
4 4 4	0	
9 8 8 7 7 5 5	0*	8
4 4 3 0 0 0	1	3
	1*	5 7 7 9
	2	0 0 0 1 1 3 3
	2*	5 8 8

eBookplus



TUTORIAL
eles-1403
Worked example 20



**THINK**

- 1 Calculate the five-number summary for Group A, in which there are 16 scores.
- 2 Calculate the five-number summary for Group B. There are 16 scores here also.
- 3 Write a five-number summary for each group.
- 4 Draw the boxplots from the five-number summaries for both groups.
- 5 Compare the data. Consider central score, highest and lowest scores, variability in scores, and so on.

WRITE/DRAW

Calculate the five-number summary for group A.

Median = average of 8th and 9th scores

$$\begin{aligned} &= \frac{2.0 + 2.0}{2} \\ &= 2.0 \end{aligned}$$

Q_1 = average of 4th and 5th scores

$$\begin{aligned} &= \frac{1.7 + 1.7}{2} \\ &= 1.7 \end{aligned}$$

Q_3 = average of 4th and 5th scores from median

$$\begin{aligned} &= \frac{2.3 + 2.3}{2} \\ &= 2.3 \end{aligned}$$

Five number summary: 0.8, 1.7, 2.0, 2.3, 2.8

Calculate the five-number summary for group B.

Median = average of 8th and 9th scores

$$\begin{aligned} &= \frac{0.8 + 0.8}{2} \\ &= 0.8 \end{aligned}$$

Q_1 = average of 4th and 5th scores

$$\begin{aligned} &= \frac{0.5 + 0.5}{2} \\ &= 0.5 \end{aligned}$$

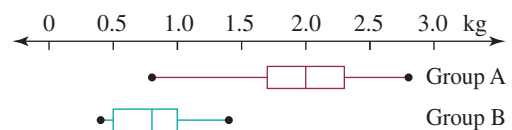
Q_3 = average of 4th and 5th scores from median

$$\begin{aligned} &= \frac{1.0 + 1.0}{2} \\ &= 1.0 \end{aligned}$$

Five number summary: 0.4, 0.5, 0.8, 1.0, 1.4

Group A: 0.8, 1.7, 2.0, 2.3, 2.8

Group B: 0.4, 0.5, 0.8, 1.0, 1.4



- The biggest of all chickens was from Group A (hormone group).
- The smallest of all chickens was from Group B (no hormone).
- The Group A data showed a little more variability (Group A interquartile range = 0.6, Group B interquartile range = 0.5).
- The median size of chickens in Group A was larger (Group A median = 2, Group B median = 0.8).
- Over three-quarters of the Group A chickens were bigger than all of the Group B chickens!

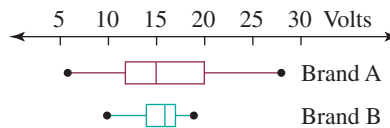
Conclusion:

The growth hormone proved to be effective.

Exercise 1H Comparing sets of data

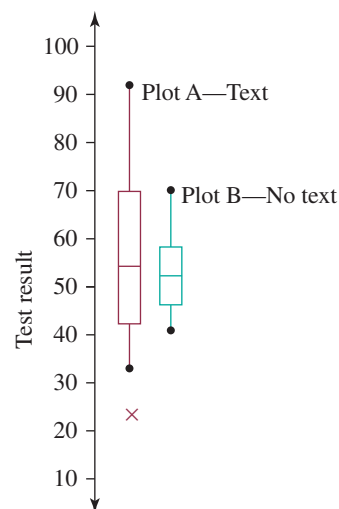
- 1 The boxplots at right show the results of testing two brands of transistor by applying increasing voltage.

- Which brand had the transistor that could withstand the highest voltage?
- Which brand had the transistor that could withstand the least voltage?
- Which brand gave the best median performance?
- Which brand showed most variability in terms of range?
- Which brand showed most variability in terms of interquartile range?
- Which brand would you recommend to manufacturers of electronic equipment if they requested a transistor that reliably worked at its expected voltage?
- Which brand would you recommend to manufacturers of electronic equipment if they requested a transistor that was likely to withstand higher voltages?



- 2 The boxplots at right were drawn by a teacher who was trying to assess the effects of allowing students the privilege of an 'open book' exam. Plot A gives information about the results of students who were allowed to use their textbook in an end of unit test. Plot B gives information about the results of students who did the test without the aid of the textbook.

- Which group had the student who gained the best test result?
- Which group did best on median result?
- Compare the variability in the performance of each group.
- Does the use of the textbook get a better test performance? Explain.
- What other things need to be taken into account when drawing these conclusions?



- 3 Draw parallel boxplots for the following pair of five-number summaries.

Group X: 14, 18.5, 21.5, 27.5, 33

Group Y: 11, 17.5, 21, 26.5, 35

- 4 **WE20** The following stem-and-leaf plot gives the age at marriage of a group of 10 women and a group of 10 men.

Key: 1 8 = 18 years old			
Men		Stem	Women
Leaf			Leaf
8 7		1	8 8
9 8 7 5 1		2	0 2 3 4 4 5
6 3		3	0 1
0		4	

- Draw parallel boxplots of the data.
 - Make comparisons about the distribution of the sets of data.
- 5 The number of words in each of the first 12 sentences is counted in each of three different types of book: a children's book, a grade 12 geography text and a major daily newspaper. The results are as follows:

Children's book	6	8	12	15	6	8	10	8	5	11	10	8
Geography text	16	18	25	13	10	25	29	18	7	22	28	22
Newspaper	12	6	8	14	18	7	12	10	21	17	16	8

- Draw parallel boxplots of the data.
- Make comparisons about the sentence length of each type of publication. Use statistics in your answer.

eBookplus

DIGITAL DOC
doc-9532
Boxplots

- 6 The stem-and-leaf plot below gives the batting scores of two cricket players — Smith and Jones — who share the responsibility of ‘opening the batting’ for their side.

Key: $1|2 = 12$

Jones Leaf	Stem	Smith Leaf
3	0	0 0 1
	1	2 6 9
8 7 4 2	2	
9 9 8 7 7 5	3	
8 4 4 2 0	4	6 6 8
5 2 0	5	7 8 8 9 9
1	6	0 4 6
	7	2 4
	8	5

- Derive a five-number summary for each player.
- Draw side-by-side boxplots of the data.
- Make comparisons between the two sets of data. Use statistics in your answer.
- Which player do you consider to be the best ‘opening batter’ and why?

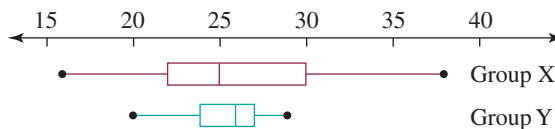
Questions 7 and 8 refer to the following stem-and-leaf plot.

Key: $12|2 = 122$

Group B Leaf	Stem	Group A Leaf
	12	2
	13	3 8
6	14	0 4 4 6
8 5 4 2 2	15	2 3 5 7 8
8 5 5 3 0 0	16	2 4 4 5
7 4 4 1 0	17	2 6
1 1	18	1

- 7 **MC** The lower quartile of Group B is:
A 156.5 **B** 144 **C** 155 **D** 152 **E** none of these
- 8 **MC** Which of the following statements is untrue?
A Data from Group A are normally distributed.
B Data from Group B are skewed.
C Group B has a greater median.
D Group A shows a greater amount of variability.
E None of the above (All of the statements are true.)

Questions 9 and 10 refer to the boxplots at right.



- 9 **MC** Which of the following statements is a correct comparison of the data?
A Group X has a higher median and shows more variability than Group Y.
B Group X has a lower median and shows more variability than Group Y.
C Group X has a higher median and shows less variability than Group Y.
D Group X has a lower median and shows less variability than Group Y.
E It is impossible to make comparisons like this without seeing the data displayed on a stem-and-leaf plot.
- 10 **MC** Which of the following statements is untrue of the boxplots?
A One-quarter of all Group Y data is less than any of Group X data.
B The median of Group X is 25.
C The interquartile range of Group X is 25.
D The range of Group Y is 9.
E None of the above (All the statements are true.)

- 11 A packing machine is meant to pack sacks of flour in 20.0 kg weights. A quality control manager notices that the machine appears to be too generous in the amount that it is putting into each sack. After checking a sample of packs for weight he adjusts the machine. After some time he selects a second sample of packs and checks their weights. The results are detailed on the stem-and-leaf plot below.

Key: 20|3 = 20.3 kg
20*|5 = 20.5 kg

After adjustment			Before adjustment	
Leaf	Stem		Leaf	
4 3	19			
9 8 8 7 7 6 5 5	19*			
4 3 1 1 0 0	20		3 4 4	
9 8 6	20*		5 5 7 8 8 8 9	
4 3 0	21		0 0 1 1 2 3 3 3 4	
6 6	21*		5 5 6 7 7 8	
2	22			

- a Derive a five-number summary for each set of data.
b Draw parallel boxplots of the data.
c Compare the performance of the machine before and after it was adjusted. Use statistics in your answer.
d Should the manager have adjusted the machine? Why (not)?
- 12 A new spray treatment has been developed to improve the budding of apple trees. Twenty-five trees are subjected to the spray treatment while 25 others are kept as a control. The number of apples that form on each tree is recorded below.



Group A (sprayed)

35 62 71 21 34 42 76 45 48 32
29 85 73 28 34 59 62 66 27 29
33 38 54 42 61

Group B (unsprayed)

44 42 65 41 39 68 63 62 58 51
43 47 49 45 40 52 56 50 71 39
35 37 38 62 68

- a Detail the data on a back-to-back stem-and-leaf plot. Use a class size of 10.
b Prepare side-by-side boxplots of the data.
c Make comparisons of the data. Use statistics in your answer.
d Comment on the overall effect of the spray. Would you recommend that orchardists use the spray?
- 13 Twenty different flashlight bulbs of each of two brands are tested until they burn out. The lifetime of each (in days) is recorded below.

Glow-worm

23 45 31 38 39 41 48 47 54 23
28 35 42 49 50 41 52 48 27 35

Starlet

28 16 24 36 47 18 59 32 64 68
72 35 46 72 54 31 29 36 55 43

- a Detail the data on a back-to-back stem-and-leaf plot. Use a class size of 5.
b Prepare side-by-side boxplots of the data.
c Make comparisons of the data. Use statistics in your answer.
d Which brand would you recommend as best? Why?

- 14** A new drug for the relief of cold symptoms has been developed. To test the drug, 40 people were exposed to a cold virus. Twenty patients were then given a dose of the drug while another 20 patients were given a placebo. (In medical tests a control group is often given a ‘placebo’ drug. The subjects in this group believe that they have been given the real drug but in fact their dose contains no drug at all.) All participants were then asked to indicate the time when they first felt relief of symptoms. The number of hours from the time the dose was administered to the time when the patients first felt relief of symptoms are detailed below.

Group A (drug)

25	29	32	45	18	21	37	42	62	13
42	38	44	42	35	47	62	17	34	32

Group B (placebo)

25	17	35	42	35	28	20	32	38	35
34	32	25	18	22	28	21	24	32	36

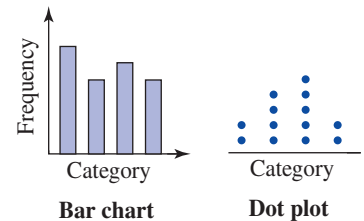
- a** Detail the data on a back-to-back stem-and-leaf plot.
- b** Prepare parallel boxplots of the data.
- c** Make comparisons of the data. Use statistics in your answer.
- d** Does the drug work? Justify your answer.
- e** What other considerations should be taken into account when trying to draw conclusions from an experiment of this type?



Summary

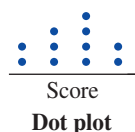
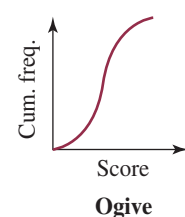
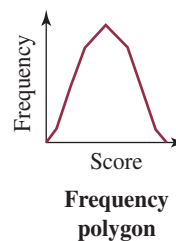
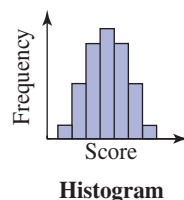
Categorical data

- All data can be classed as *categorical* (the data fit a qualitative category) or *numerical* (each observation is a quantitative amount).
- Categorical data may be *nominal* (no order attached to the categories) or *ordinal* (the categories can be placed in a definite order).
- Numerical data can be either *discrete* (take only certain fixed values) or *continuous* (may take any value within a range).
- A *frequency distribution table* shows the number of times each category of data occurs.
- The *frequency* of an observation is the number of times that it occurs.
- The *relative frequency* of an observation is the number of times that it occurs as fraction of the total number of pieces of data collected.
- The *percentage frequency* is the relative frequency expressed as a percentage.
- Categorical data may be displayed on a *bar chart* or a *dot plot*.



Numerical data

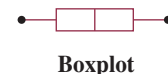
- *Numerical data* are formed when each record is a quantitative amount and can be placed on a numerical scale. Numerical data can be either:
 1. *discrete* (take only certain fixed values) or
 2. *continuous* (may take any value within a range).
- A piece of numerical data is known as a *score*.
- Numerical data may be displayed on a *histogram*, *polygon*, *ogive*, *dot plot*, *stem-and-leaf plot* or *boxplot*.



Key: 2|6 = 26

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

Stem-and-leaf plot



- A *histogram* is similar to a bar graph but has no gaps between the columns, and frequency is always plotted on the vertical axis.
- A *frequency polygon* is a line graph, obtained by joining the centres of the tops of each column of the histogram.
- Histograms and polygons give an indication of the distribution of the data. The patterns of the distribution are normal, skewed (to the left or to the right), bimodal, clustered or spread.

Cumulative data

- The *cumulative frequency* of a score is the number of times that score, plus all lower scores, occurs in the set of data.
- An *ogive* is a line graph of the cumulative frequency results.
- A *percentile* is the score below which a particular percentage of the data lies.

Measures of central tendency

- The centre of a set of scores can be defined by the *mean*, *median* or *mode*.
- The *mean* of a set of scores is the average of the scores and is given by $\bar{x} = \frac{\sum x_i}{n}$.
- The *median* of a set of scores is the middle score when the scores are placed in order of size. The median is the $\frac{n+1}{2}$ th score.
- The *mode* is the most frequently occurring score.
- If the data contain extreme values (outliers), the median (or mode) is a better measure of central tendency than the mean.

Measures of variability

- The *spread* or *variability* of a set of scores can be measured by finding the *range*, the *interquartile range*, the *standard deviation* or the *variance*.
- The *range* of a set of scores is the difference between the highest and lowest scores:

$$\text{Range} = X_{\max} - X_{\min}.$$

- The *lower quartile* Q_1 is the score which occurs one-quarter of the way through the distribution. It can be found by dividing the distribution in half, then taking the median of the lower half.
- The *upper quartile* Q_3 is the score which occurs three-quarters of the way through the distribution. It can be found by dividing the distribution in half, then taking the median of the upper half.
- The *interquartile range* is the difference between the upper and lower quartiles:

$$IQR = Q_3 - Q_1.$$

- The *standard deviation* of a group of scores can be found using the formula:

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

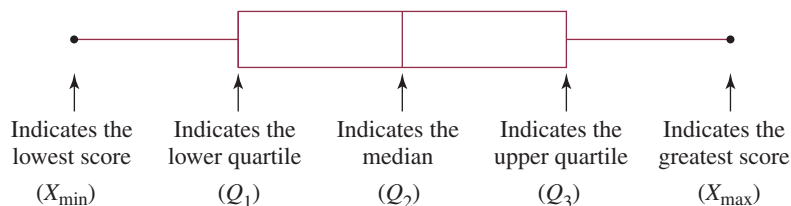
- The *variance* is the square of the standard deviation: $\text{variance} = s^2$.

Stem-and-leaf plots

- Stem-and-leaf plots are used to group and summarise data.
- Every plot should include a key; the leaves should appear in ascending order.
- To use a class size of 5, use * notation on class stems.
- The plot gives a graphical representation of the spread of the data.
- The quartiles and the median can be easily obtained from the plot.

Boxplots

- A *five-number summary* is a list, consisting of the lowest score ($X_{\min}$), lower quartile (Q_1), median (Q_2), upper quartile (Q_3) and the greatest score ($X_{\max}$) (in that order) of the data.
- A *boxplot* is a graph of the five-number summary.
- The boxplot is a powerful tool to show the spread of the data.
- Boxplots are always drawn to scale.
- The box spans the interquartile range; the median is marked by the vertical line inside the box; the whiskers extend to the lowest and greatest scores.



- The extreme values can be denoted by a small cross; the whiskers then are shortened to the next largest (or smallest) value.

Comparing sets of data

- *Back-to-back* stem-and-leaf plots:
 1. are useful to compare the distribution of two similar sets of data
 2. share the same stem
 3. contain a key, which usually relates to the data on the right
 4. are arranged such that the data on the left increase outwards from the plot.
- *Parallel* or *side-by-side* boxplots:
 1. are useful for quantitative comparisons
 2. share a common scale
 3. compare two or more sets of data.

Chapter review

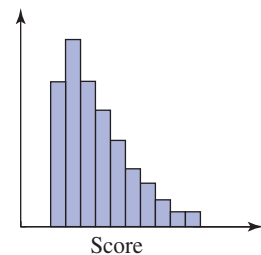
MULTIPLE CHOICE

- 1 The time taken for each competitor in the interschool crosscountry event is recorded. The data formed could be described as:
- A nominal data B ordinal data C discrete data
D continuous data E categorical data
- 2 The country of birth of every student is recorded on their enrolment form. The data formed could be described as:
- A nominal data B ordinal data C discrete data
D continuous data E numerical data

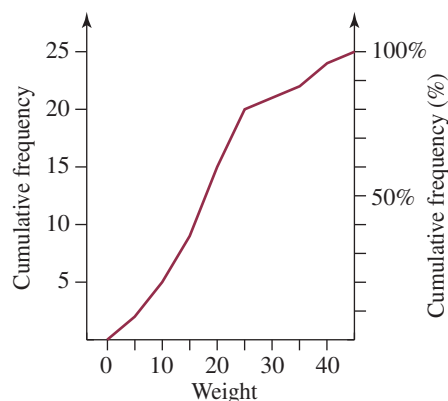
Questions 3 and 4 refer to the following frequency distribution table.

Response	Frequency
Often	2
Sometimes	6
Occasionally	7
Rarely	4
Never	1

- 3 The data represented in the table could be referred to as:
- A nominal data B ordinal data C discrete data
D continuous data E numerical data
- 4 The relative frequency of the response 'rarely' is:
- A 4 B $\frac{4}{20}$ C 20%
D 4% E none of these
- 5 The distribution of data shown by the histogram below could be described as:
- A clustered data
B spread data
C positively skewed data
D negatively skewed data
E bimodal data



Questions 6 and 7 refer to the ogive below.



- 6 How many of the items had a weight of 20 kg or less?
- A 5 B 10 C 15 D 20 E 25
- 7 The 90th percentile of the data is:
- A 6 B 40 C 22.5 D 36 E none of these

- 8 Find the median of the following data.

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

A 1.5 B 12 C 6.5 D 4.5 E 4

- 9 Find the median of the following data.

Score	Frequency
15	12
16	15
17	10
18	6
19	3
20	2

A 24.5 B 15 C 16 D 17.5 E 16.56

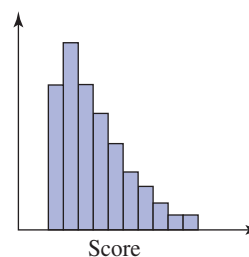
- 10 Find the mean of the following data.

Score	Frequency
20–	3
30–	6
40–	12
50–	13
60–	7
70–	2

A 44.88 B 50 C 45
D 49.88 E None of the above

- 11 Which of the following would be most likely to be true of the data represented by this histogram?

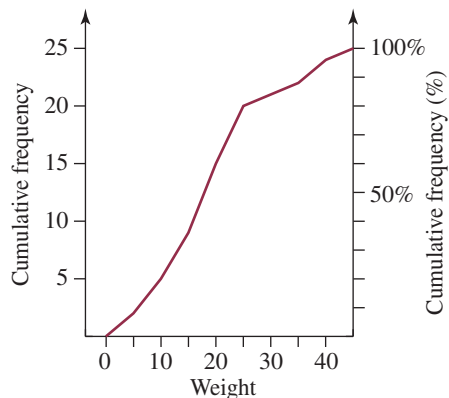
A The mode would exceed the mean.
B The mean would exceed the mode.
C The mean would equal the mode.
D The lower quartile would equal the median.
E The interquartile range would exceed the range.



- 12 Which of the following is not generally affected by the inclusion of an extreme value in a set of data?

A The range B The mean C The interquartile range
D The standard deviation E The variance

- 13 The interquartile range of the data shown below is:



A 45 B 17 C 24 D 12 to 24 E 12

14 Which of the following is not a measure of the variability of a set of scores?

- A The range
- B The mean
- C The interquartile range
- D The standard deviation
- E The variance

15 The median of a set of data is the 13th score. The position of the lower quartile will be:

- A the 6th score
- B the 7th score
- C the 6.5th score
- D the 19th score
- E indeterminable without seeing the data itself

Questions 16 and 17 refer to the following boxplot.



16 The upper quartile of the data is:

- A 14
- B 24
- C 28
- D 36
- E 42

17 The interquartile range is:

- A 28
- B 12
- C 24 to 36
- D 14 to 42
- E 42

18 Find the interquartile range of the data represented by the following stem-and-leaf plot.

Key $10|3 = 103$

Stem	Leaf
10	2 3
11	4 5 6 8
12	0 0 3 5 9
13	2 3 6 7 8
14	1 1 2
15	3

- A 20.5
- B 2.5
- C 2
- D 205
- E 20

19 Find the median of the data, represented by the following stem-and-leaf plot.

Key $6|2 = 6.2$

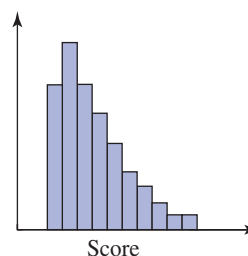
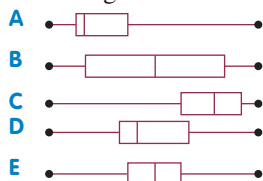
Stem	Leaf
6*	2
6	5 6 9
7	1 1 2 3 3 4
7*	5 5 6 7 8 8 9
8	0 0 1 4
8*	7 9

- A 75
- B 75.5
- C 7.55
- D 7.5
- E 0.75

20 Which of the following cannot be easily determined from a boxplot?

- A The range
- B The median
- C The upper quartile
- D The standard deviation
- E The interquartile range

- 21 Which of the following boxplots would best represent the data shown in this histogram?



- 22 The number 36 was obtained using this five-number summary: 28, 36, 50, 62, 64.
The number 36 must be:

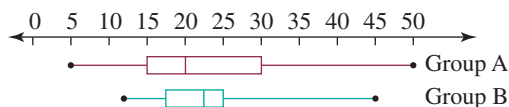
- A the mean
B the median
C the range
D the IQR
E the upper quartile

- 23 Which of the statements above right is true for this back-to-back stem-and-leaf plot?

Key: 23|0 = 230

Group B	Stem	Group A
Leaf		Leaf
3 2 1	23	0
7 7 7 6 5 4 0	24	1 5
8 6 5 2 2	25	0 2 5 8
7 4	26	3 5 6 6 8
9	27	4 7 8
6	28	2 6
2	29	0 1

- A Data from Group A is bimodal.
B Data from Group B is normally distributed.
C Group A has the larger IQR.
D Group B has the larger range.
E Group A has the smaller median.
- 24 Which of the statements below is true of these boxplots?



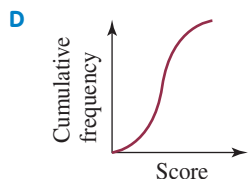
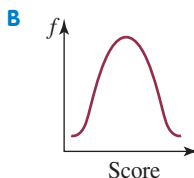
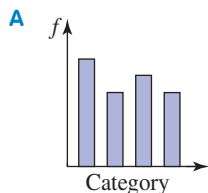
- A Group A has higher median and shows more variability than Group B.
B Group B has higher median and shows more variability than Group A.
C Data from Group A are spread.
D Data from Group B are positively skewed.
E One-quarter of all Group A data is more than any of Group B data.

- 1 For each of the following, indicate the type of data that would be formed. Your answer should be one of the following: nominal, ordinal, discrete or continuous.

- a The time taken for each member of the class to get to school
b The number of goals scored each week by a soccer team
c People are asked which of five designs they like as a new flag.
d The annual income of each teacher in the school
e The number of pages in each novel submitted for this year's Miles Franklin Award
f People are asked to judge their level of contentment with life and are given a number of alternative responses with which they can answer.

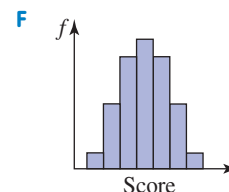
SHORT
ANSWER

2 Match each of the following graph types with the correct graph.



E

Key 2 6 = 26	
Stem	Leaf
2	6 9
3	1 4 5 6 9
4	1 2 2 8
5	0 2



- | | |
|--------------------|-----------------------------|
| a Boxplot | b Stem-and-leaf plot |
| c Dot plot | d Bar chart |
| e Histogram | f Frequency polygon |
| g Ogive | |

3 Thirty customers at a supermarket were asked to take part in a taste test of a new product. Each was asked to rate the product by choosing one of the words: poor (P), fair (F), good (G), very good (V) or excellent (E). The results are shown below.

G	G	V	G	E	F	P	G	V	V
V	E	G	G	G	F	G	P	V	V
E	V	V	F	F	G	V	G	G	V

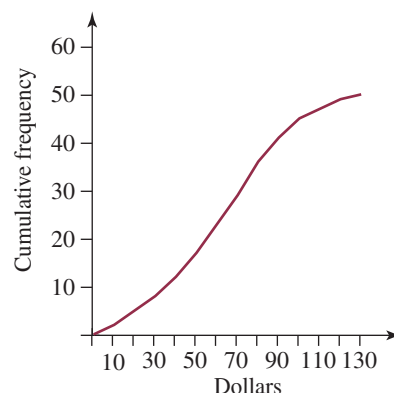
- Display the data by using a dot plot.
 - Find the frequency of a response of good or better.
 - Find the relative frequency of a response of good or better.
 - Find the percentage frequency of a response of good or better.
- 4 The following data give the number of maths subjects, taken by 30 students over a 2-year period (Years 11 and 12).

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

- Represent the data on a frequency distribution table.
- Draw a histogram for the data.
- Add a polygon to the histogram.
- What was the relative frequency of two maths subjects?
- What was the percentage frequency of three maths subjects?

5 A researcher investigates the amount of money spent by 50 people in a local Coles supermarket on Friday night between 6 pm and 8 pm. His findings are presented in the ogive shown at right.

- How many people spent \$50 or less?
- How many people spent more than \$100 on shopping?
- What percentage of people spent less than \$80?
- Find the 80th percentile of the data and write a short sentence explaining what it means.
- If Nathan was in the lowest 10% of buyers, what is the maximum amount of money he could have spent?



- 6 The following scores were recorded: 2, 6, 9, 1, 3, 5, 6, 7, 8, 2. For these data find:
- a the mean b the median c the mode
d the range e the interquartile range f the standard deviation
g the variance.

7

Score	Frequency
30–	3
40–	6
50–	12
60–	15
70–	18
80–	10

For the above set of data find:

- a the mean b the median group
c the modal group d the range
e the standard deviation f the variance.
- 8 The following data give the amount of cut meat (in kg) obtained from each of 20 lambs.
- 4.5 6.2 5.8 4.7 4.0 3.9 6.2 6.8 5.5 6.1
5.9 5.8 5.0 4.3 4.0 4.6 4.8 5.3 4.2 4.8
- a Detail the data on a stem-and-leaf plot. (Use a class size of 0.5 kg.)
b Prepare a five-number summary of the data.
c Draw a boxplot of the data.
- 9 The side-by-side stem-and-leaf plot below shows the number of lessons per week given by two piano teachers, Victoria and Elena, over the period of 30 consecutive weeks.

Key: 1|3 = 13 lessons

Elena	Stem	Victoria
Leaf		Leaf
4 3 2 2	0	3 4
9 8 8 7 6 6 5 5	0*	6 8 9
4 4 3 2 2 2 1 0 0	1	0 2 4 4
7 6 6 5 5	1*	5 5 7 8 9
2 2 1	2	0 0 1 2 3 4
5	2*	5 5 6 7 9
	3	3 4
	3*	5 5

- a Derive a five-number summary for each set of data.
b Draw a side-by-side boxplot of the data.
c Make comparisons between the two sets of data. Use statistics in your answer.

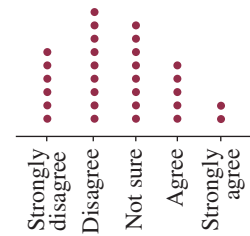
- 1 In preparation for an upcoming debate, Rachel, a debating team captain, decided to conduct a small survey. She asked a number of students whether they agreed with the following statement: 'LOTE should be a compulsory subject in Years 7 to 10'. Students were given a choice of the following options:

- strongly disagree
- disagree
- not sure
- agree
- strongly agree.

EXTENDED
RESPONSE

Rachel recorded the responses using a dot plot shown at right.

- How many people participated in the survey?
- State whether the data obtained in the survey are categorical or numerical.
- Classify the data further by selecting an appropriate word from this list: continuous, nominal, discrete, ordinal.
- Represent the information by using a bar chart.
- Construct a frequency distribution table for the data.
- What was the relative frequency of students who agreed with the proposed statement?
- Find the percentage frequency of students who were not sure about their response.
- A particular response option was selected by 20% of all students. What was that option?



- 2 The following data give the marks of 30 students in a physics test.

90	88	72	45	32	63	72	65	64	69
61	58	58	59	72	93	95	21	34	56
58	69	62	64	52	58	57	44	72	63

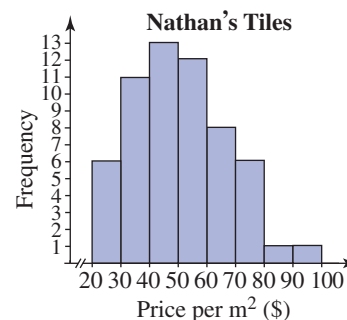
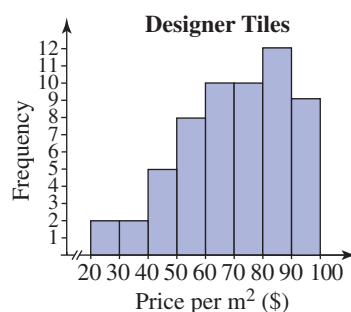
- Tally the data on a frequency distribution table. Use a class size of 10.
- Prepare a histogram of the data.
- Prepare an ogive of the data.
- Find the median, upper quartile and lower quartile from the ogive.
- The top 10% of students are to be given a certificate of excellence. What mark would be needed to get a certificate?
- Draw a box-and-whisker diagram of the data.

- 3 The following data show the weekly spending (in dollars) of 25 Year-12 students on entertainment.

10	5	0	2	4
5	22	18	6	8
2	0	0	5	12
16	18	2	4	8
24	17	8	6	8

- Tally the data on a frequency distribution table. Use a class size of 5.
- Prepare a histogram of the data.
- Find the mean of the data.
- Find the median class of the data.
- Find the modal class of the data.
- Find the standard deviation of the data.
- Find the variance of the data.

- 4 Nathan owns two shops: Designer Tiles and Nathan's Tiles. Although the shops are stocked identically, he notices that in Designer Tiles he mostly sells expensive porcelain tiles, while in Nathan's Tiles the demand for cheaper ceramic tiles seems to be greater. To see the trends more clearly, Nathan uses his sales records to construct the histograms as shown below.



- a For each shop describe the distribution of the data, shown by the histograms.
- b Use the histograms to arrange the data for each shop into frequency distribution tables.
- c Perform all necessary calculations and complete the following table.

	Statistics	Designer Tiles	Nathan's Tiles
i	Mean		
ii	Median class		
iii	Modal class		
iv	Range		
v	Standard deviation		

- d Use the statistics obtained in part c to compare the buying trends in the two shops.
 - e Nathan feels that he needs to rearrange the stock in his shops. He plans to increase the range of cheaper ceramic tiles in Nathan's Tiles. He also wants to move most of his expensive porcelain stock to Designer Tiles. Would you support Nathan's decision? Justify your answer.
- 5 In order to compare two textbooks a teacher recommends one book to one of his classes and another book to another class. At the end of the year the classes are each tested. The results are detailed below:

Text A (25 students)

44 52 95 76 13 94 83 72 55 81
 22 25 64 72 35 48 56 59 84 98
 84 21 35 69 28

Text B (28 students)

65 72 48 63 68 59 68 62 75 79
 81 72 64 53 58 59 64 66 68 42
 37 39 55 58 52 82 79 55

- a Prepare a back-to-back stem-and-leaf plot of the data.
- b Prepare a five-figure summary for each group. (Note that the groups are different sizes.)
- c Prepare parallel boxplots of the data.
- d Compare the performance of each of the classes.
- e Which textbook do you think would be best? Why?
- f What other things would you need to take into account before drawing final conclusions?

eBookplus

DIGITAL DOC
 doc-9833
 Test Yourself
 Chapter 1

Chapter opener**DIGITAL DOC** doc-9523

- 10 Quick Questions: Warm up with ten quick questions on univariate data. (page 1)

1A **Categorical data****DIGITAL DOCS**

- doc-9524: Investigate graph types. (page 4)
- SkillSHEET 1.1 doc-9525: Practise converting a fraction to a decimal or a decimal to a percentage. (page 4)
- doc-9526: Investigate dot plots. (page 5)

1B **Numerical data****TUTORIAL**

- **WE5** eles-1337: Watch how to represent univariate data on a frequency distribution table and to draw a histogram and a frequency polygon. (page 10)

DIGITAL DOCS

- doc-9527: Investigate one-variable statistics. (page 11)
- doc-9528: Investigate one-variable statistics — grouped. (page 11)

1C **Cumulative data****TUTORIAL**

- **WE6** eles-1338: After constructing an ogive, watch how to use it to determine the 75th and 90th percentiles of data provided in a frequency table. (page 15)

DIGITAL DOCS

- doc-9529: Investigate one-variable statistics. (page 16)
- WorkSHEET 1.1 doc-9530: Represent univariate data graphically, interpret and classify univariate data. (page 19)

1D **Measures of central tendency****TUTORIAL**

- **WE9** eles-1339: Watch how to find the mean, median and modal class of data provided in a frequency table. (page 22)

DIGITAL DOC

- doc-9527: Investigate one-variable statistics. (page 23)

1E **Measures of variability****TUTORIALS**

- **WE12** eles-1340: By finding the 25th and 75th percentiles, watch how to determine the interquartile range of data provided in a frequency table. (page 28)
- **WE14** eles-1341: Watch how to determine the mean, standard deviation and variance of data provided in a frequency table using a CAS calculator. (page 30)

DIGITAL DOC

- WorkSHEET 1.2 doc-9531: Determine summary statistics and construct frequency histograms and ogives. (page 33)

1G **Boxplots****INTERACTIVITY** int-0802

- Boxplot and five-number summary: Consolidate your understanding of boxplots. (page 37)

DIGITAL DOC

- doc-9532: Investigate boxplots. (page 41)

1H **Comparing sets of data****TUTORIAL**

- **WE20** eles-1403: Use a CAS calculator to construct parallel boxplots from a back-to-back stem-and-leaf plot. (page 43)

DIGITAL DOC

- doc-9532: Investigate boxplots. (page 45)

Chapter review**DIGITAL DOC** doc-9833

- Test Yourself Chapter 1: Take the end-of-chapter test to test your progress. (page 59)

To access eBookPLUS activities, log on to www.jacplus.com.au

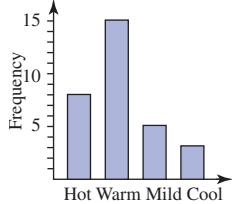
Answers CHAPTER 1

UNIVARIATE DATA

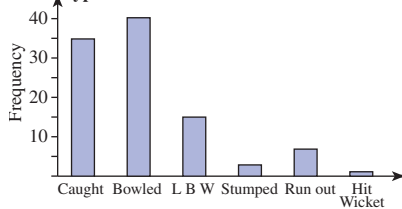
Exercise 1A Categorical data

- 1 a Numerical b Categorical
c Numerical d Categorical
e Numerical
- 2 a Ordinal b Nominal
c Nominal d Ordinal
e Ordinal
- 3 a Continuous b Discrete
c Continuous d Continuous
e Discrete
- 4 B 5 D

6 Weather type during January

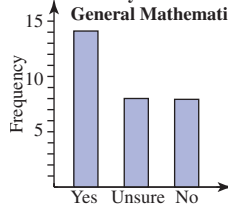


7 Type of dismissal



- 8 a Frequencies: Yes = 14, Unsure = 8, No = 8

b Would you still choose General Mathematics?



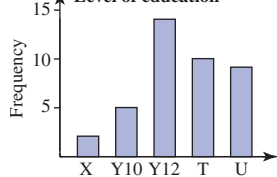
- c 14
d $\frac{7}{15}$
e 46.67%
f Ordinal, categorical data

9 C

- 10 a Married b 15
c $\frac{3}{8}$ d 37.5%
e 25 f 62.5%

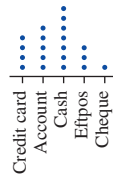
- 11 a Frequencies: X = 2, Y10 = 5, Y12 = 14, T = 10, U = 9

b Level of education



- c 14
d 19
e 47.5%
f Ordinal, categorical data

12 a



- b 7 c 35%
d Nominal, categorical data

13 C

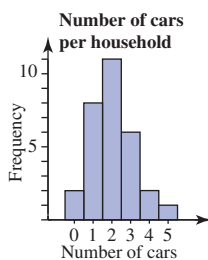
14 a



- b Good c 90

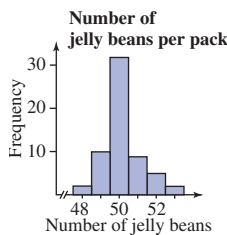
Exercise 1B Numerical data

1 a



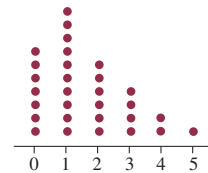
- b 6 c $\frac{1}{5}$
d 20%

2 a



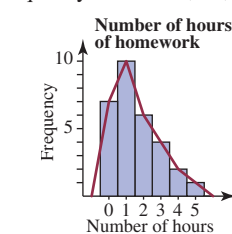
- b 32 c $\frac{32}{60} = \frac{8}{15}$
d 53.33%

3 a



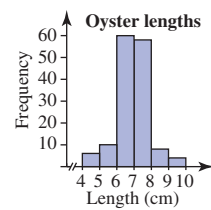
- b Frequency column: 7, 10, 6, 4, 2, 1

c, d



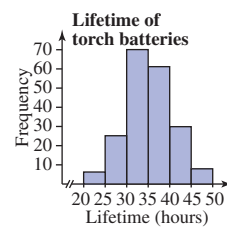
- e 1 f 6
g $\frac{6}{30} = \frac{1}{5}$ h 20%

4 a



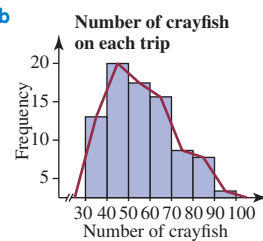
- b $10, \frac{10}{146} = \frac{5}{73}, 6.85\%$

5 a



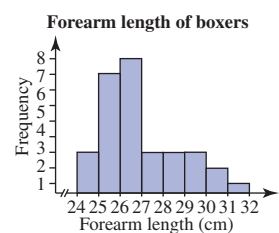
- b $70, \frac{70}{200} = \frac{7}{20}, 35\%$

6 a, b



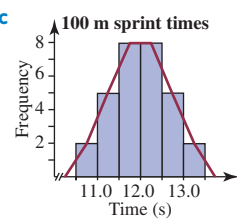
- c Positively skewed data
7 a Frequency column: 3, 7, 8, 3, 3, 3, 2, 1

b

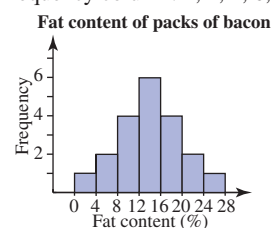


- c Positively skewed data
8 a Frequency column: 2, 5, 8, 8, 5, 2

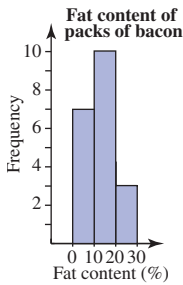
b, c



- d Normally distributed data
9 a Frequency column: 1, 2, 4, 6, 4, 2, 1



- b Frequency column: 7, 10, 3

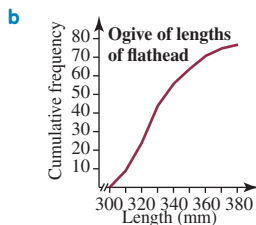


- c The data appear more clustered if the group size is larger.

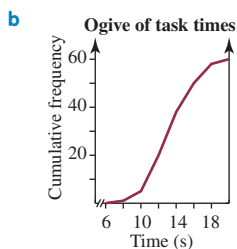
10 E
11 C
12 C
13 D

Exercise 1C Cumulative data

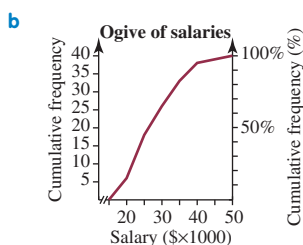
- 1 a Cumulative frequency column: 9, 24, 44, 56, 64, 71, 75, 77



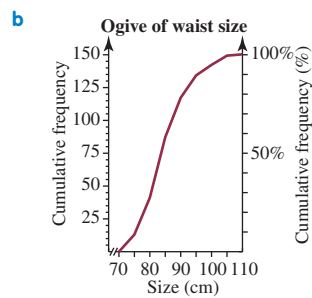
- 2 a Cumulative frequency column: 1, 5, 20, 38, 50, 58, 60



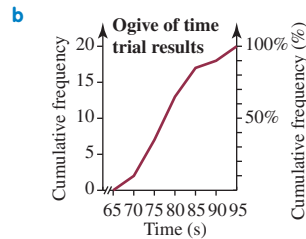
- 3 a Cumulative frequency column: 6, 18, 26, 33, 38, 39, 40



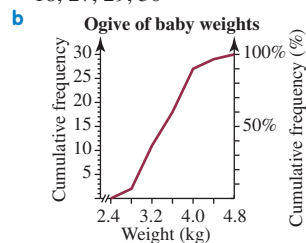
- c 11
d 27
e \$33 000
f \$26 500
g \$21 500
h \$38 000
4 a Cumulative frequency column: 13, 41, 87, 117, 134, 142, 149, 150



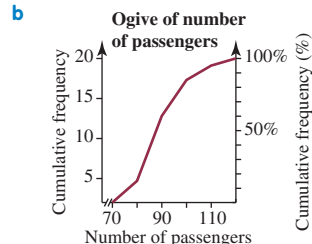
- c 60
e 96 cm
g 23%
5 a 12
c 68%
e 55
g 37
d 20
f 84 cm
b 11
d 8%
f 46
h 30 or less
6 a Cumulative frequency column: 2, 7, 13, 17, 18, 20



- c 14
d 90% of the riders finished with a time of 90 s or less.
e To qualify you need a time under 72 s.
7 a Cumulative frequency column: 2, 11, 18, 27, 29, 30

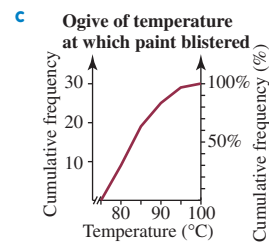
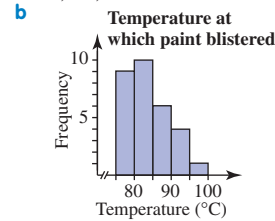


- c 75%
d 3.45 kg
e 3%
8 a Cumulative frequency column: 3, 12, 17, 19, 20



- c 13
d 97 passengers
e 89 passengers
f 81 passengers
g No

- 9 a Cumulative frequency column: 9, 19, 25, 29, 30



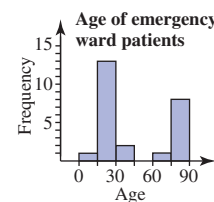
- d 19
f 26%
h 78 °C
e 63%
g 93 °C

10 D

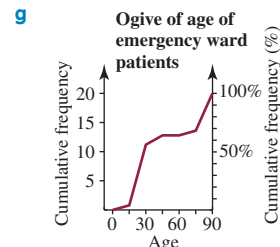
11 C

Exercise 1D Measures of central tendency

- 1 a Mean = 6.286, median = 6, mode = 6
b Mean = 102.9, median = 102, mode = 102
c Mean = 5.425, median = 5.4, mode = 5.4
d Mean = 27.4, median = 27.5, mode = 27 and 28
2 a 1.58
b 1
c 1
3 a 12.01
b 10–14
c 0–4 and 5–9
4 a 152.59
b 150–<200
c 150–<200
5 B
6 B
7 E
8 a Frequency column: 16, 6, 4, 2, 1, 1
b 7.33
c 0–5 hours
d 0–5 hours
9 a Frequency column: 1, 13, 2, 0, 1, 8
b



- c Bimodal
e 16–30
d 44.1
f 16–30

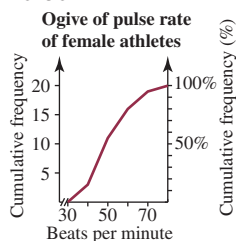


h 30

i No

- 10 a Player A mean = 34.33,
Player B mean = 41.83
b Player B
c Player A median = 32.5, Player B
median = 0
d Player A

- 11 a Frequency column: 3, 8, 5, 3, 1
b 50.5 c 40–50
d 40–50
e



f 49

- 12 A 13 B 14 A

15 Various answers

16 Various answers

Exercise 1E Measures of variability

- 1 a 5 b 9
c 1.0 d 6.94
2 E 3 35
4 a 5 b 9
c 2 d 4
5 22 6 0.8 7 C
8 a 2.446 b 2.370
c 21.567 d 3.427
9 As for Question 8.

10 D

11 C

- 12 a Group A: mean = median =
mode = 170 cm
Group B: mean = median =
mode = 170 cm
b No c Group B
d Group B e Group B
f Group A: range = 20, interquartile
range = 0, $s = 5.774$
Group B: range = 120, interquartile
range = 20, $s = 35.119$

13 E

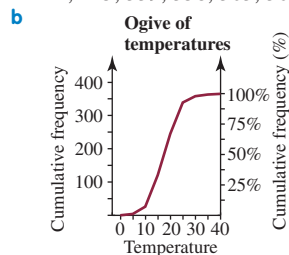
- 14 a 60 b 110.48
c $s = 11.79$, $s^2 = 139.00$

15 C

- 16 a Mean = 455.29, $s = 88.88$
b $s^2 = 7899.65$

17 B

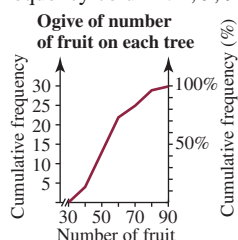
- 18 a Cumulative frequency column: 4, 26,
121, 245, 339, 358, 363, 365



- c $Q_3 = 21$, $Q_1 = 13.5$, interquartile
range = 7.5

- d 17 e 17.55
f 5.657 g 32.002
h 40

- 19 a Frequency column: 4, 9, 9, 3, 4, 1



- c Median = 55, $Q_3 = 63$, $Q_1 = 45$
d 18 e 54
f 13.48 g 45

- 20 a Electric Mate

b Electric Mate = 197 hours, Hot
Wire = 185 hours

c Electric Mate

d, e Various answers

- 21 Crinkle would be the best buy because the
spread of packet weight is closer to the
mean than Crunch.

Exercise 1F Stem-and-leaf plots

- 1 Key: 0|6 = 6

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

- 2 Key: 3|6 = 36 min

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

- 3 Key: 8*|6 = 86 wpm

Stem	Leaf
8*	6 8
9	2
9*	5 5 5 6 6 6 8 9
10	2 2 2 3
10*	7 7 7 8 8 8
11	0 1 2
11*	5
12	0 1 1 4
12*	0
13	

- 4 Key: 13*|8 = 13.8

Stem	Leaf
13*	8 9
14	0 2 3 3
14*	5 6 6 7 7 8 8
15	1 2 2
15*	5 6 7 9

- 5 1.95 g

- 6 a 8.4 s

b $Q_3 = 8.85$ s, $Q_1 = 7.8$ s

c 1.05 s

7 A

8 A

9 B

10 D

- 11 a Key: 20|2 = 20.2 cm

Stem	Leaf
20	2 7
21	4 6 7 8
22	2 8 8 8
23	0 1 2 6 6
24	6 6 7 8
25	1

- b 22.9

c $Q_3 = 24.1$, $Q_1 = 21.75$

d 2.35

- 12 a Key: 8|2 = 82 cm

Stem	Leaf
8	2 4 5 5 8
9	0 1 2 3 4 6 6 7 8 9
10	1 1 3 3 7 8
11	0 0 1 4 5 7
12	6 6
13	2

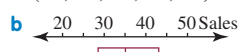
- b 100 cm

c $Q_3 = 110$ cm, $Q_1 = 92$ cm

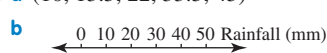
d 18 cm

Exercise 1G Boxplots

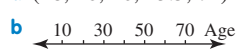
- 1 a 13 b 5 c 26
2 a 122 b 6 c 27
3 a 49.0 b 5.8 c 18.6
4 a 140 b 56 c 90
d 84 e 26
5 a 58 b 31 c 43
d 27 e 7
6 C 7 C 8 D
9 a (22, 28, 35, 43, 48)



- 10 a (10, 13.5, 22, 33.5, 45)

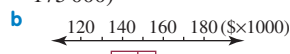


- 11 a (18, 20, 26, 43.5, 74)



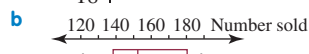
c Positively skewed including one
extremely high value

- 12 a (124 000, 135 000, 148 000, 157 000,
175 000)



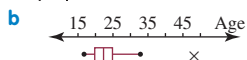
- 13 a Key: 12|1 = 121

Stem	Leaf
12	1 5 6 9
13	1 2 4
14	3 4 8 8
15	0 2 2 2 5 7
16	3 5
17	2 9
18	1 1 1 2 3 7 8



- 14 a Key: 1*|7 = 17 years

Stem	Leaf
1*	7 7 8 8 8 9 9
2	0 0 0 1 2 2 2 2 3 3 3 3 4 4 4
2*	5 5 8 9
3	1 2 3
3*	
4	
4*	8



- c Positively skewed with one high value. Most mothers have their babies between the ages of 20 and 25.

15 C

Exercise 1H Comparing sets of data

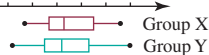
- 1 a Brand A b Brand A c Brand B
d Brand A e Brand A f Brand B
g Brand A

- 2 a Text b Text

- c The 'text' group showed more variability.

- d and e Various answers

- 3 10 15 20 25 30 35 Scale

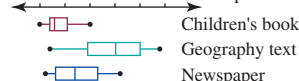


- 4 a 15 20 25 30 35 40 Age



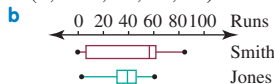
- b Women had a smaller range of 13 compared with men, who had a range of 23. The IQR for women was 5, whereas for men it was 12. The median age of women was 23.5 and for men it was 27.5. There was less variability in the marriage age for women.

- 5 a 5 10 15 20 25 30 Words per sentence



- b The longest sentence was in the geography book (29), then newspaper (21) then children's book (15). The shortest sentence was in the children's book (5), then newspaper (6) then geography book (7). The geography book had the largest median (20), then the newspaper (12) then the children's book (8).

- 6 a Smith: (0, 17.5, 57.5, 62, 85), Jones: (3, 31.5, 39, 46, 61)



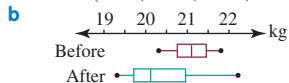
7 C

8 E

9 B

10 C

- 11 a Before: (20.3, 20.75, 21.1, 21.45, 21.8)
After: (19.3, 19.7, 20.1, 20.95, 22.2)

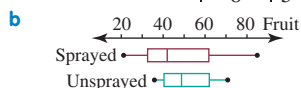


- c The consistency decreased because the results became more variable. Higher IQR of 1.25 kg versus 0.7 kg and overall range of 2.9 versus 1.5.

- d The machine should be adjusted.

- 12 a Key: 2|1 = 21 fruit

Unsprayed		Sprayed	
Leaf	Stem	Leaf	Stem
9 9 8 7 5	2	1 7 8 9 9	1
9 7 5 4 3 2 1 0	3	2 3 4 4 5 8	2
8 6 2 1 0	4	2 2 5 8	3
8 8 5 3 2 2	5	4 9	4
1	6	1 2 2 6	5
	7	1 3 6	6
	8	5	7

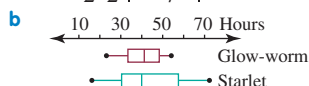


- c Sprayed fruit has a greater range (64) and IQR (29.5), showing variability in the results. The median is lower. Unsprayed fruit has a more consistent result with a smaller range (36) and IQR (21.5). The median is higher.

- d It would probably be better not to spray as the smaller spread provides better 'control' over the size of the harvest.

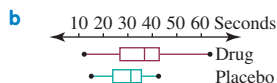
- 13 a Key: 2|3 = 23 hours

Starlet		Glow-worm	
Leaf	Stem	Leaf	Stem
8 6	1*		
4	2	3 3	3
9 8	2*	7 8	4
2 1	3	1	5
6 6 5	3*	5 5 8 9	6
3	4	1 1 2	7
7 6	4*	5 7 8 8 9	8
5 4	5	0 2 4	9
9	5*		
4	6		
8	6*		
2 2	7		



- 14 a Key: 1|3 = 13 seconds

Placebo		Drug	
Leaf	Stem	Leaf	Stem
8 8 5 5 4 2 1 0	1	3 7 8	3
8 6 5 5 5 4 2 2 2	2	1 5 9	4
	3	2 2 4 5 7 8	5
	4	2 2 2 4 5 7	6
	5		
	6	2 2	7



- c to e Variability versus Consistency. The placebo group showed a more consistent performance with a better median time – 30 seconds versus 36 seconds, a smaller range – 25 versus 49, and a smaller IQR = 12 versus 16. Even though some people

felt relief quicker in the drug group – 4 seconds faster on minimum, the drug group results varied enormously. Therefore the drug did not seem to enhance the group's performance on a whole.

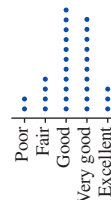
CHAPTER REVIEW

MULTIPLE CHOICE

- 1 D 2 A 3 B
4 B 5 C 6 C
7 D 8 D 9 C
10 D 11 B 12 C
13 E 14 B 15 C
16 D 17 B 18 A
19 D 20 D 21 A
22 C 23 C 24 D

SHORT ANSWER

- 1 a Continuous b Discrete
c Nominal d Discrete
e Discrete f Ordinal
2 a G b E
c C d A
e F f B
g D
3 a



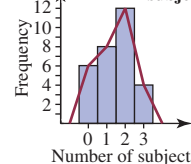
- b 24

$$c \frac{24}{30} = \frac{4}{5}$$

- d 80%

Number of subjects	Frequency
0	6
1	8
2	12
3	4

b, c Histogram and polygon of the number of maths subjects



- d $\frac{2}{5}$ e $13\frac{1}{3}\%$

- 5 a 17

- b 5

- c 72%

- d \$88; 80% of people spent \$88 or less on shopping.

- e \$20

- 6 a 4.9

- c 2 and 6

- e 5

- g 7.656

- 7 a 65.781

- c 70–

- e 13.836

- b 5.5

- d 8

- f 2.767

- b 60–

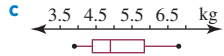
- d 60

- f 191.443

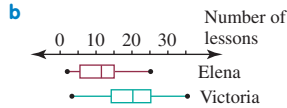
- 8 a Key: $3^*|9 = 3.9$

Stem	Leaf
3*	9
4	0 0 2 3
4*	5 6 7 8 8
5	0 3
5*	5 8 8 9
6	1 2 2
6*	8

- b (3.9, 4.4, 4.9, 5.85, 6.8)



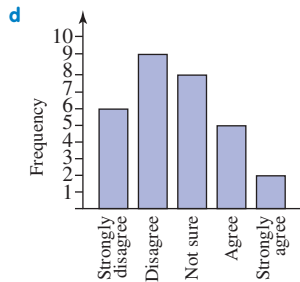
- 9 a Elena: 2, 6, 11.5, 15, 25;
Victoria: 3, 14, 20, 25, 35



- c The number of lessons given by Victoria has varied compared with Elena. Elena's number has been more consistent.

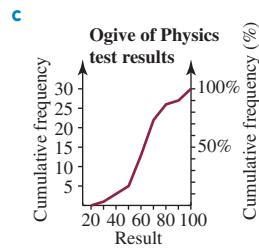
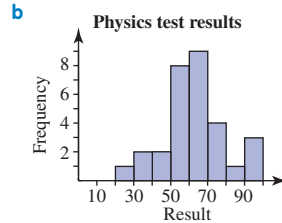
EXTENDED RESPONSE

- 1 a 30 b Categorical
c Ordinal

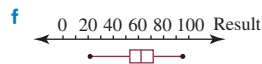


- e Frequency column: 6, 9, 8, 5, 2
f $\frac{1}{6}$
g $26\frac{2}{3}\%$
h 'Strongly disagree'

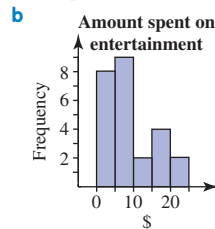
- 2 a Frequency column: 1, 2, 2, 8, 9, 4, 1, 3



- d Median = 62,
 $Q_3 = 72$
 $Q_1 = 53$
e 90



- 3 a Frequency column: 8, 9, 2, 4, 2



- c 9.1
d $5.00 < 10.00$
e $5.00 < 10.00$
f 6.570
g 43.17
4 a Designer Tiles: negatively skewed;
Nathan's Tiles: positively skewed

- b Frequency column:

Designer Tiles: 2, 2, 5, 8, 10, 10, 12, 9
Nathan's Tiles: 6, 11, 13, 12, 8, 6, 1, 1

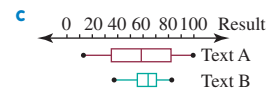
	Statistics	Designer Tiles	Nathan's Tiles
i	Mean	70	50.52
ii	Median class	70–<80	40–<50
iii	Modal class	80–<90	40–<50
iv	Range	80	80
v	Standard deviation	18.85	16.56

- d Measures of central tendency are all higher for Designer Tiles. Spread is almost the same with standard deviation for Designer Tiles slightly higher. Overall trends: in Designer Tiles people tend to buy more expensive tiles.
e Yes, as this will be in accordance with buying trends.

- 5 a Key: $1|3 = 13$

Text B		Stem	Text A
Leaf			Leaf
		1	3
		2	1 2 5 8
	9 7	3	5 5
	8 2	4	4 8
9 9 8 8 5 5 3 2		5	2 5 6 9
8 8 8 6 5 4 4 3 2		6	4 9
	9 9 5 2 2	7	2 2 6
	2 1	8	1 3 4 4
		9	4 5 8

- b Text A: (13, 35, 59, 82, 98);
Text B: (37, 55, 63.5, 70, 82)



- d to f Various answers

