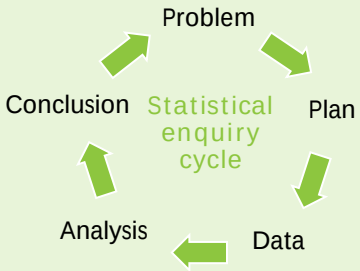
Mathematics and Statistics 1.12

Externally assessed
4 credits

Demonstrate understanding of chance and data

The statistical enquiry cycle (PPDAC)

The **statistical enquiry cycle** summarises the steps involved in a statistical investigation.



Statistical literacy involves the ability to understand, interpret and evaluate the results of statistical investigations undertaken by others.

Data collection

It is often impractical to carry out a **census** (an investigation involving every member of a **population** of interest). Instead, a portion of the population, known as a **sample**, is investigated.

- The sample should be **random** (each member of the population has the same chance of being chosen) so that the characteristics of the sample are similar to those of the population.
- If the sample does not accurately reflect the characteristics of the whole population then the sample is said to be **biased**.

Statistical analysis of a random sample allows **inferences** (conclusions) to be made about the population as a whole.

In a statistical investigation, data may be collected by the investigators themselves. Alternatively, a data set may be obtained from a secondary source – it is important that this **data source** is listed.

Selecting a random sample

Suppose you are given a list of 500 students, and you have to select a random sample of 30 names from the list.

'Drawing names from a hat' is a good reliable method that has no bias.

A quicker method is to use the random numbers on your calculator, as follows:

- give each student a number from 1 to 500
- set your scientific calculator to produce random numbers from 1 to 500
  (Press: 5 0 0 Ran# + 1 = and take the whole number part, ignoring repetitions)
- obtain the first thirty random numbers
- identify the sample of 30 students by matching the numbers to the names.

Example

Select a random sample of 8 names from the following list of 15 names:

John, Will, Anne, Helen, Henry, Tom, Liz, Luke, Nathan, Jacob, Angus, Amy, Jack, Sally, Pat

Solution

Enter the names in a table and give a number to each name.

Number	Name
1	John
2	Will
✔ 3	Anne
✔ 4	Helen
5	Henry
6	Tom
✔ 7	Liz
✔ 8	Luke

Number	Name
✔ 9	Nathan
10	Jacob
✔ 11	Angus
12	Amy
✔ 13	Jack
✔ 14	Sally
15	Pat

Using your calculator, obtain 8 numbers in the range 1–15

(Press: 1 5 Ran# + 1 = and take the whole number part only, ignoring repetitions)

A typical result might be:

13, 3, 4, 8, 13 discard (no repeats), 7, 14, 11, 9

Highlight or tick the names, as shown.

Ans. p. 51

In a statistical investigation it is important that the data collection method is verified as **random**, so that the sample is representative of the population (shares its features). This allows **inferences** to be made about the population.

Exercise A: Selecting a random sample

Use the random numbers on your calculator to answer the questions in Exercise A.

1. Select 8 names from the list of 12 names in the table.

Number	1	2	3	4	5	6	7	8	9	10	11	12
Name	Sarah	Millie	Phiz	Boz	Liz	John	Bill	Ann	James	Barry	Amy	Petra

List the eight numbers, and in the table highlight your choices.

Random numbers used:

2. Allocate a number to each country in the table. Select 10 countries from the list of 15. Highlight your selection and list the numbers.

Number																
Country	Brazil	Peru	Bolivia	Sudan	Libya	Korea	Angola	Zambia	USA	NZ	UK	China	Niger	Poland	Kuwait	

Random numbers used:

3. At Mountain High School there is a boys' soccer team and a girls' netball team. The players in each squad are listed in the table. Select a sample of 7 players from the soccer team and 5 players from the netball team.

Soccer team																
Number																
Name	Luke	Peter	John	Ian	Carl	Andy	Sean	Karl	Mike	Kurt	Amos	Sang	Mark	Joel	Alan	

Random numbers used:

Netball team													
Number													
Name	Grace	Meg	Emma	Ella	Lorna	Sarah	Prue	Emily	India	Rose	Laura	Bianca	

Random numbers used:

Data collection methods

There are three main methods of gathering sample data.

- **Observation** – watching and accurately recording the information, e.g. counting cars (**discrete** data), measuring plant heights (**continuous** data), etc.
- Oral **interviews** – an investigator asks questions and records responses, e.g. opinions about products, or attitudes to various issues (**qualitative** data).
- Written **questionnaires** – also involve questions and responses, but in written form. Answers are recorded in various ways – in words, choosing from multichoice options, or using a scale (e.g. a rating from 0 to 10).

The initial data gathered is called **raw data**.

Data organisation

After collection, the raw data are often organised into **tables**. This allows features of the data set to be seen more clearly.

Frequency distribution tables

An effective way to organise raw data is the **frequency distribution table**. If there is a small number of data values, then scores are listed separately.

Example

The times taken in seconds for a sample of 5-year-old children to complete a simple puzzle are shown below:

6, 5, 5, 7, 7, 6, 5, 9, 7, 6, 6, 6, 8,
6, 5, 9, 5, 6, 9, 5, 6, 5, 7, 8, 6

The data are organised, using tallies, into a frequency table.

Time (seconds)	Tally	Frequency
5	𝍸 ||	7
6	𝍸 ||||	9
7	||||	4
8	||	2
9	|||	3
Total		25

9 children took 6 seconds.

5 children took more than 7 seconds.

Frequency tables for grouped data

For **discrete** (counted) data sets with many values, the data are grouped into **classes**, e.g. 1–5.

Continuous (measured) data are always grouped into intervals of values, since there are infinitely many values possible.

Example

The distance students travelled to school each day was measured (rounded to the nearest tenth of a kilometre). The results for one class are shown.

2.3, 1.8, 0.8, 2.4, 2.1, 3.0, 0.7, 0.5, 1.1, 1.8, 1.4, 2.7, 3.1, 1.1, 0.9, 2.0, 0.2, 1.4, 1.0, 3.2, 4.4, 1.8, 1.5, 0.7, 0.1, 1.6, 2.3, 2.2, 1.7, 0.3

The data are organised into classes of width 0.5 km.

Distance (km)	Tally	Frequency
0.0–	𝍸	3
0.5–	𝍸	5
1.0–	𝍸	5
1.5–	𝍸 |	6
2.0–	𝍸 |	6
2.5–	|	1
3.0–	𝍸	3
3.5–		0
4.0–4.5	|	1
		30

The table clearly shows that one distance is much greater than the others (4.4 km).

Note: The frequency table is not unique – other interval widths are possible.

Continuous data cannot be listed, and are always grouped in classes.

Example

The time taken to travel to school one day was recorded in a table by each student in a class.

Time taken (minutes)	Tally	Number of students
0–	𝍸 ||||	9
15–	𝍸 𝍸 𝍸 ||	17
30–	||||	4
45–60	|	1

The following observations can be made.

1. In the table, the interval 0– means

   0 minutes ≤ time taken < 15 minutes.
2. If a student took 32 minutes to travel to school, a **tally** would be placed in the 30– class.
3. 4 + 1 = 5 students took more than 30 minutes to travel to school.
4. No students took longer than 1 hour to travel to school.
5. If a student is randomly selected from the class, the time the student took to get to school would most likely be between 15 and 30 minutes, since this is the class with the highest frequency.

Exercise B: Frequency distribution tables

Ans. p. 51

1. As part of a survey, students record the number of pieces of fruit they ate the previous weekend. The results are shown for one class:

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

   a. Present these data in a frequency table.

   b. Comment on any interesting features of the data that can be seen.

c. Would the above data be useful for answering the question, 'Does being home at the weekend influence the amount of fruit students eat per day?' Comment on at least two aspects.

2. Junior cyclists at a school are timed over an 8-km course. Their times (in minutes rounded to 2 d.p.) are as follows:

19.21	17.65	17.23	19.28	17.01	23.45
22.11	23.08	21.01	19.97	17.65	18.34
16.56	18.57	23.62	14.76	19.36	18.89
15.90	21.04	19.34	18.52	16.77	19.43

The data are presented in a grouped frequency table (using interval widths of 1 minute).

Time (min)	Tally	Frequency
14–	|	1
15–	|	1
16–	||	2
17–	||||	4
18–	||||	4
19–	𝍸 |	6
20–		0
21–	||	2
22–	|	1
23–24	|||	3
	Total	24

Comment on any interesting features of the data that can be seen.

3. Millie weighed a random sample of school bags from 50 Year 11 students. Her results are in the table alongside.

Weight (kg)	Number of students
0–	9
2–	18
4–	12
6–	8
8–10	3
	50

a. Explain what the interval 4– means.

b. Which frequency would change if another bag of weight 5.2 kg is entered into the table?

c. Millie says, "Over half of the bags in the sample weigh more than 5 kg". Is this claim supported by the data in the table?

d. A bag was left behind after the survey was done. What would be the most likely range of weights for the bag?

e. Millie says, "Most Year 11 students' bags weigh less than 6 kg". Comment on her statement, discussing what factors would make her claim likely to be true or false.

Analysing data

Sample statistics such as averages and measures of spread are usually calculated in a statistical investigation. These statistics summarise the patterns in the data set.

Measures of centre

A **statistical average**, is a measure of central tendency, or 'middle' value of a set of data. The three types of **average** are:

- **Mean** = $\frac{\text{total of all scores}}{\text{number of scores}}$
- **Median** – the middle score (or average of the two middle scores) when the data are **ranked** (put in order of size).
- **Mode** – the most frequently occurring score in the data set. (There may be two modes.)

Example

For the following marks out of ten in a spelling test:

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

Mean = $\frac{\text{total of scores}}{\text{number of scores}}$

$= \frac{100}{20}$ [the 20 scores add up to 100]

$= 5$

Ranking the scores:

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

↑ median

The median is halfway between the middle two scores, which are both 5

The median is 5

The mode is 6 since 6 is the most common score.

Graphics calculators can be used to calculate measures of centre automatically. Learn how your calculator does this.

Comparing measures of centre

Each type of average has strengths and weaknesses.

- The mean uses all data values but can take on values which are not typical of the data set. This often occurs because of extreme values (which are very high or low when compared with the rest of the data).
- The median is unaffected by outliers, but gives no indication that they exist.
- The mode is easy to calculate, but is of limited significance. Also, the value of the mode can change a lot as new scores are added to a data set. If there are more than two scores with the same highest frequency then the mode does not exist at all for that data set.

Exercise C: Measures of centre

Ans. p. 51

1. The money spent (to the nearest dollar) during a visit to a town's Easter Show was recorded for a group of fifteen high school students.

12	29	32	26	18	22	27	33
17	16	27	21	27	18	20	

   a. Find the median amount spent.

   b. Find the mean amount spent.

   c. Find the mode for this set of data.

2. The numbers of goals scored during the season by each member of a soccer team were recorded and listed in order:

0 0 1 2 4 5 6 7 9 9 12 16 19 25

   a. Find the mean number of goals scored per team member.

   b. i. Complete the statement:

   Half the team scored _____ goals or fewer.

   ii. What is the median for the data set?

   c. i. How many numbers could be used to complete the statement: John and Jaime both scored … goals?

   ii. Comment on the mode of this data set.

Ans. p. 51

Averages from frequency tables

The following example illustrates how to find the mean for data in a frequency table.

Example

Thirty customers were asked how many times they had visited the supermarket in the previous week. The frequency table shows the results.

Number of visits (x)	Frequency (f)	fx
0	1	0
1	6	6
2	6	12
3	8	24
4	6	24
5	3	15
Totals	30	81

The mean number of visits = $\frac{81}{30}$

[the total of the fx values is the sum of all scores]

= 2.7

The median number of visits is the number of visits halfway between the 15th and 16th visit (when numbers of visits are put in order).

Adding the frequencies, it can be seen that
1 + 6 + 6 = 13 visited 2 times or fewer and
1 + 6 + 6 + 8 = 21 visited 3 times or fewer. So the median must be 3 (15th and 16th score both equal 3).

For grouped discrete data or continuous data, the midpoint of the interval, m, is used instead of exact x-values. fm-values are then calculated and totalled (instead of fx-values). This allows an estimate of the mean to be calculated.

Example

The times taken (in seconds) for a group of fourteen children to complete a simple puzzle are shown in the table below.

x	f	m	fm
0–	2	5	10
10–	3	15	45
20–	7	25	175
30–40	2	35	70
Totals	14		300

[the midpoint, m, of the 30–40 interval is the middle value, 35, etc.]

The mean is estimated as $\frac{300}{14}$ = 21.4 seconds (1 d.p.)

Note: An interval such as 0– means

0 ≤ time taken < 10, etc.

Exercise D: Averages from frequency tables

1. The table below shows the number of plastic bags per customer used by thirty-six consecutive customers at a minimart.

Number of bags (x)	Number of customers (f)	
0	1	
1	3	
2	5	
3	7	
4	6	
5	8	
6	4	
7	2	
Totals		

a. Complete the table in order to find the mean number of bags used per customer.

b. i. What is the mode for this data set?

ii. Explain the meaning of this mode in everyday language.

c. Find the median number of bags used.

d. Which of the three averages would be used in a press release which claimed, "50% of our customers use fewer than … bags per visit".

2. Vicky recorded the times taken by the fastest twenty students in her class to complete a number puzzle, 'Studoku'. The table shows her data:

Time (min) *x*	Frequency *f*	Midpoint *m*	*fm*
5–	1	6	6
7–	3	8	
9–	8		
11–	5		
13–15	3		
Totals			

a. Complete the table and use it to estimate the mean time taken by the fastest twenty students to complete 'Studoku'.

b. The slowest 10 students in Vicky's class took an average of 21 minutes to complete the puzzle. Find the overall average time for Vicky's class of 30 students.

3. Marks in a science project were grouped to give the grades A, B and C, as shown in the table.

Grade	Mark *x*	Frequency *f*
A	20–24	3
B	15–19	5
C	10–14	12

Estimate the mean mark and grade.

4. James has an average of 72% in his first four exams. If he averages 75% in his five exams he has been promised a CD that he wants. What mark does he need in his last exam in order to get this reward?

Quartiles and measures of spread

The **upper quartile** (UQ) of a set of scores is the median of the upper half of the scores.

The **lower quartile** (LQ) of a set of scores is the median of the lower half of the scores.

Example

27 scores are randomly selected from a vocabulary test marked out of ten. The scores are listed in order:

3, 4, 4, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, 10, 10, 10, 10

⬆ lower quartile ⬆ median ⬆ upper quartile

There are 27 scores so the median is the 14th score which is 7.

There are 13 scores in the lower half of the data (up to but not including the median), so the median of the lower half is the 7th score which is 6.

Similarly, the upper quartile is the 21st score which is 8 (the median is also excluded from the upper half).

Note: The medians and the quartiles divide the sample into quarters: the proportion of the data lying below the LQ is $\frac{1}{4}$ (or 25%). Similarly, $\frac{1}{2}$ (or 50%) of the data is below the median and $\frac{3}{4}$ (or 75%) of the data lies below the UQ.

The **range** of a sample is a measure of the spread of the data.

range = highest score – lowest score

In the vocabulary test sample above the range is 10 – 3 = 7

The **interquartile range** (IQR) of a sample is the range of the middle half of the data.

IQR = upper quartile – lower quartile

In the vocabulary test sample above the interquartile range is 8 – 6 = 2

Ans. p. 52

Exercise E: Quartiles and measures of spread

1. The money spent (to the nearest dollar) during a visit to a town's Easter Show was recorded for a group of fifteen high school students.

12 29 32 26 18 22 27 33
17 16 27 21 27 18 20

a. For the amount spent, find:

i. the lower quartile

ii. the upper quartile

b. What is the interquartile range?

2. The numbers of goals scored during the season by each member of a sample of soccer players were recorded and listed in order:

0 0 1 2 4 5 6 7 9 9 12 16 19 25

For the number of goals scored in a season by players in this sample, find:

a. the upper quartile

b. the lower quartile

c. the interquartile range

3. The numbers of students in a random sample of Year 7 classes in a city were:

23 24 24 25 26 28 28 29 31 32

a. Find the range of the sample.

b. Find the lower quartile.

c. Find the upper quartile.

d. Find the interquartile range.

4. Find the LQ, median and UQ for the following samples.

a. Dollar spend at supermarket:
23, 35, 67, 89, 92, 99, 100, 125, 130

b. Cost of loaf of bread:
$2.45, $3.65, $3.80, $3.90, $4.20, $4.50

c. Numbers per hour entering pool:
22, 25, 16, 19, 32, 21, 18, 45, 20, 16, 15

d. Time (minutes) to complete puzzle:
7.6, 3.4, 9.1, 8.4, 6.3, 7.6, 5.1, 10.3, 4.7

5. Find the range and interquartile range for each of the data sets in question 4.

a. Range =

IQR =

b. Range =

IQR =

c. Range =

IQR =

d. Range =

IQR =

Answers

Exercise A: Selecting a random sample (page 2)

1. – 3. *Answers will vary*.

Exercise B: Frequency distribution tables (page 3)

1. a.

No. of pieces	Tally	Frequency
0	𝍸	2
1	𝍸 |	6
2	|||	3
3	𝍸 |	6
4	𝍸 ||	7
5	𝍸	5
6	|	1
Total		30

b. *Answers will vary*, e.g. the most common number of pieces of fruit was 4. Only two people had no fruit all weekend.

c. No, since students may not have been at home. Also there is no data available about fruit consumption during the week for comparison.

2. *Answers will vary*, e.g. most riders had times between 17–19 min. Only six riders took more than 20 minutes.

3. a. 4 kg ≤ weight < 6 kg

b. The frequency for 4– would increase from 12 to 13

c. This is incorrect since more than half the bags (27 bags out of 50) weigh less than 4 kg

d. Between 2 kg and 4 kg

e. Since the sample is random and of size fifty, this statement is likely to be true, since 78% of the bags in the sample are below 6 kg (so around 20% of bags are heavier than 6 kg). However, if the sample is not representative of the population of all Y11 bags, then the statement may be untrue.

Exercise C: Measures of centre (page 5)

1. a. $22 b. $23 c. $27

2. a. 8.21 (3 s.f.)

b. i. 6 ii. 6.5

c. i. Two ii. Two modes: 0 and 9

Exercise D: Averages from frequency tables (page 6)

1.

Number of bags *x*	Number of customers *f*	Number of bags *fx*
0	1	0
1	3	3
2	5	10
3	7	21
4	6	24
5	8	40
6	4	24
7	2	14
Totals	36	136

a. $3\frac{7}{9}$

b. i. 5

ii. Five was the most common number of bags used by a customer

c. 4 d. Median

2.

Time (min) *x*	Frequency *f*	Midpoint *m*	*fm*
5–	1	6	6
7–	3	8	24
9–	8	10	80
11–	5	12	60
13–15	3	14	42
Totals	20		212

a. 10.6 minutes b. 14.1 minutes

3. 15 (nearest whole number) B

4. 87%

Exercise E: Quartiles and measures of spread (page 8)

1. a. i. $18 ii. $27

   b. $9

2. a. 12 b. 2 c. 10

3. a. 9 b. 24 c. 29

   d. 5

		LQ	Median	UQ
4.	a.	$51	$92	$112.50
	b.	$3.65	$3.85	$4.20
	c.	16	20	25
	d.	4.9 min	7.6 min	8.75 min

5. a. Range = $107 IQR = $61.50

   b. Range = $2.05 IQR = $0.55

   c. Range = 30 IQR = 9

   d. Range = 6.9 min IQR = 3.85 min

Exercise F: Dot plots and box-and-whisker plots (page 10)

Comments will vary – some suggested answers are given

1. The distribution of the sample of Year 11 boys' neck circumferences is approximately symmetrical and bell-shaped about the median neck circumference of 35.5 cm .

   One neck circumference measurement was not supplied for the sample (n = 29).

   There is one unusually large neck circumference of 60 cm, and one unusually small neck circumference of 16 cm, both of which are likely to be errors (students supply their own measurements for the data base).

   The middle 50% of male neck circumferences, as shown by the box, are very consistent, with an interquartile range of 4.3 cm.

2. The distribution of the sample of Year 11 girls' heights is symmetrical and approximately rectangular in shape, so as a result the mean and the median are very similar in value (about 165 cm).

   One height measurement was not supplied for the sample (n = 29).

   There are no very large or small heights in the sample, though one Year 11 girl is 180 cm, which is about 5 cm taller than the next tallest girl in the sample, but not an unlikely height.

   The middle 50% of female heights, are symmetrically distributed, with an interquartile range of 8 cm.

3. The distribution of the sample of Year 11 boys' memory game times is skewed right (the lowest half of times were from 27 sec to 42.5 sec but the upper half of times were from 42.5 sec to 94 sec). As a result there is a difference of more than 5 seconds between the mean game time (47.8 sec) and the median game time (42.5 sec) – the mean memory game time was pulled up by a few extra long times.

   Most memory game times for these Year 11 boys are between 30 sec and 60 sec, but a few were longer than this. The maximum game time of 94 seconds seems unusually long (more than double the median time) – this could be because the boy lost focus during the test, or didn't follow the instructions correctly.

   The middle 50% of game times are also skewed right, with an interquartile range of 18.75 sec.

4. The distribution of the sample of Year 11 girls' bag weights is a little skewed to the right (the lowest half of bags weighed between 0.5 kg and 4 kg, while the upper half of bag weights weighed between 4 kg and 10.5 kg). As a result the mean (4.41 kg) is 0.41 kg heavier than the median bag weight. The distribution is bimodal, with a cluster of bag weights around the minimum weight and another cluster around the median bag weight.

   Most bag weights for these Year 11 girls are between 0.5 kg and 6 kg, but a few were

INDEX