Mathematics and Statistics 1.10

Internally assessed
4 credits

AS 91035

Investigate a given multivariate data set using the statistical enquiry cycle

Multivariate data sets

A **variable** is a characteristic that can take on a range of values, such as a person's *height* or the *colour* of a house or the *number of pets* in a household.

A **multivariate data set** has several variables for each member of the set. For example, a multivariate data set may involve a sample of people. For each person values may be given for variables such as: gender, age, colour of eyes, height, weight, length of right foot, writing hand preference, resting heart rate, and so on.

A few lines of a typical multivariate data set are shown.

Person	1	2	3
Gender	Male	Female	Female
Age	14	15	13
Colour of eyes	Blue	Brown	Green
Height (cm)	168	153	144
Weight (kg)	61	57	48
Length of right foot (cm)	29	24	22
Writing hand	right	left	right
Resting heart rate (beats per minute)	69	71	64

In this standard you will be given a data set to investigate. This data set will usually be a **random sample** from a larger **population** (the whole group of interest).

You will be required to make a **comparison** in which the *same* variable is compared over *two different groups* in the data set.

Note: The variable you choose for the comparison needs to be one that takes on a range of numerical values (e.g. height or resting heart rate).

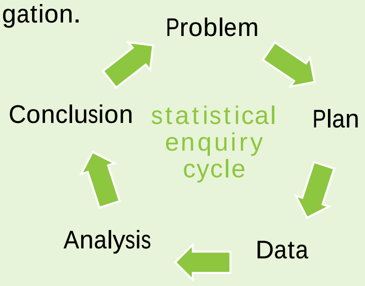
For example, you may wonder if resting heart rates (the number of heartbeats per minute for someone who is resting) are different for boys compared with girls.

- The *variable* being compared is the resting heart rate (in beats per minute).
- The two different *groups* are boys and girls (you will need to say which age groups are to be involved in your study).
- The *population* is all boys and girls in the **database** from which the sample was selected (e.g. the New Zealand CensusAtSchool data base, which is a large data bank of information for many different variables for primary and secondary school children).

The **comparison question** for your investigation will need to identify the variable, the two groups and the population.

Statistical enquiry cycle

The **statistical enquiry cycle** (**PPDAC**) is the process that should be followed during a statistical investigation.



Problem

The first stage involves defining the problem of interest, then posing a comparison question.

For example, 'I wonder if Year 11 girls tend to send more text messages than Year 11 boys'.

© ESA Publications (NZ) Ltd, Freephone 0800-372 266, ISBN 978-1-927297-63-6

AS 91035

Ans. p. 45

You should give some thought as to what answer *you* would predict for your question.

The **variables** and **terms** in **statistical questions** need to be carefully defined and understood so that planning and subsequent data collection are carried out correctly.

Plan

The following aspects should be considered when planning.

- What **data** should be collected – how much is needed and how will it be collected?
- What measurement system will be used for the variables?
- How will the data be recorded (e.g. tally charts, questionnaires)?
- What questions should be asked?
- What problems could arise in the collection of the data?

Data

Raw data are collected according to the plan. The management of the data may involve various decisions:

- Does the data need sorting?
- Are there problems/errors in the data – does the data need cleaning?

Analysis

The **analysis** stage involves summarising and making sense of the data, so that you will be able to answer your question. This may involve the following:

- Displaying the data in tables and graphs.
- Calculating statistical measures which summarise the data.
- Observing and comparing any patterns or special features of the data sets.

Are your predictions still the same, or would you now change them?

Conclusion

The question posed in the problem section is answered, with justification, based on the analysed data.

- Making an informed inference about populations using your sample data.

Any other important aspects of the investigation should also be discussed in your conclusion.

Exercise A: Statistical variables and questions

A nationwide online survey for Year 5–13 students which provides real, relevant data and classroom activities to enhance statistical enquiry across the curriculum.

CensusAtSchool www.censusatschool.org.nz is funded by *Statistics New Zealand* and the *Ministry of Education*.

The following are some of the questions that are used by CensusAtSchool in its survey.

a. Are you male or female?

b. How old are you?

c. Which country were you born in?

d. Which ethnic group do you belong to?

e. How tall are you?

f. What is the length of your right foot?

g. What is your arm span?

h. What is the circumference of your wrist?

i. What is your popliteal* length?

j. What is the length of your index finger?

k. Are you left handed, right handed or ambidextrous?

l. What is your main method of transport to school?

m. How long does it usually take you to get to school?

n. What is the weight of your school bag today?

o. What is your favourite learning area?

p. How fast is your reaction time?

q. What sport or activity do you most enjoy?

r. How physically fit do you think you are?

s. What is your resting pulse rate?

t. How long have you had your current cell phone for?

u. How many texts did you send in the previous day?

*For a seated person, the popliteal length is the measurement from the underside of the leg right behind the knee to the floor.

© ESA Publications (NZ) Ltd, Freephone 0800-372 266, ISBN 978-1-927297-63-6

1. For the above list, write down the variable involved in each question, and its unit or possible values (where appropriate).

a.

b.

c.

d.

e.

f.

g.

h.

i.

j.

k.

l.

m.

n.

o.

p.

q.

r.

s.

t.

u.

2. Suggest four suitable comparison questions using the variables in the above list. In each question, fully describe the groups being compared.

a.

b.

c.

d.

3. Which of the following would be suitable comparison questions for this standard? Tick the letters of those that are suitable. For each suitable comparison question, describe the variable being compared, its units and the two groups in the population over which it is being compared.

   For each unsuitable question, explain why it is unsuitable as a comparison question.

   a. Are 10-year-old girls generally shorter than 10-year-old boys in New Zealand?

   b. Do Year 6 boys tend to have longer arm spans than Year 6 girls in New Zealand?

   c. Is the median time taken by students who walk to school longer than the median time taken by students who travel by bus in New Zealand?

   d. Does a student's index finger tend to be shorter than their ring finger in New Zealand?

   e. Are Year 9–11 students who walk to school generally fitter than Year 9–11 students who don't walk to school in New Zealand?

   f. Are 13-year-old boys more untidy than 13-year-old girls in New Zealand?

   g. Do New Zealand students with smaller feet tend to have shorter fingers?

   h. Would Year 10 boys have a lower resting pulse rate than Year 10 girls in New Zealand?

   i. Do girls with longer arm spans tend to be taller, in New Zealand?

   j. How does the number of texts sent daily by Year 11 girls compare with the number of texts sent daily by Year 11 boys, in New Zealand?

© ESA Publications (NZ) Ltd, Freephone 0800-372 266, ISBN 978-1-927297-63-6

Data

Collecting data

In this standard, a multivariate data set will be provided.

Ideally this will be a **sample** of at least 30, selected using random (unbiased) sampling methods, so that the sample is likely to have features that are typical of the population from which it was selected.

- The bigger the sample size, the more likely it is that the sample is representative of the population from which it was selected. This makes inferences about the population more likely to be reliable.
- If sample sizes are small, it means that any conclusions you make about the population, based on the features of the sample, are likely to be invalid.

You should always think about where data came from (the data source), data collection methods, and whether the data is likely to be reliable.

An excellent source of student-related data is available at the New Zealand CensusAtSchool website. However, the data on this website is supplied by the students themselves, so may have gaps or inconsistencies that require 'cleaning' (correction or removal) before proceeding with any analysis.

Cleaning data

Tables are a useful way of recording and displaying data. All data collected should be checked for inaccuracies or errors. These could include:

- Recording errors – e.g. entering a height as 1.63 when the heights are to be recorded in centimetres. If it is simply a case of the wrong unit (metres) being used, it would be reasonable to change the value to 163 centimetres.
- Nonsense answers – e.g. 'Lunch' being recorded as a student's favourite subject. In this case the data would be ignored (and one would be very suspicious about any other data from that source).

- Missing data – e.g. no height recorded for a selected person. If you are studying heights, then unless you can go back and get this information (which is usually impossible if the data set is a sample from a database) this person must be removed from the investigation.

Example

Q. The following extract from a table of multivariate data is provided for a school database in 2014.

Student number	Sex	Year at school	Height (cm)	Bag weight (kg)	Time spent on homework the previous night (hours)
1	Male	1998	173	3.4	1.5
2		11.5	350	100	007
3	Female	10	154	2.65	2.25
4	Female	12	166	1250	1.2
5	Male	11	1.8		2.7
6	Female	13	1550	2.8	60

A. Each row of the table of the multivariate data table should be inspected for errors or inconsistencies.

Student 1 is likely to have recorded his year of birth, rather than his year at school. This would mean he was 16 and likely to be in Year 11. The rest of his data seems sensible, so a simple check could be carried out to confirm his year at school, and the entry corrected.

Student 2 has omitted information and the rest of his (or her) data is obviously incorrect (an attempt at humour?) so delete this row of data.

Student 3 seems to have supplied reasonable and complete data.

Student 4 has probably recorded her bag weight in grams: changing the bag weight to 1.25 kg seems a reasonable step to clean the data (the rest of the data seems valid).

Student 5 has omitted the bag weight but the rest of the informations seems valid, so retain this student in the data base. He has recorded his height in metres, so make the height 180 cm.

Student 6 is likely to have used mm; make the height 155 cm.

If **'dirty' data** is used in a statistical investigation, then no matter how carefully graphs are drawn or how precisely any analysis is carried out, the conclusions from the investigation will be suspect. Even one extreme value can have a big effect on the range or the mean of a data set.

Ans. p. 45

Exercise B: Cleaning data

1. Inspect the data in each row of the following spreadsheet of data gathered from students. Comment on any entries that need improvement or 'cleaning', and discuss what could be done with those entries that need 'cleaning'.

	Sex	Height (cm)	Wrist circum (cm)	Time to school (min)	Time to run 400 m (sec)
1	Male	163	16	20	68
2	Female	157	16	45	1.15
3	Boy	1.7	13	30	68
4	Girl	154	8.3	15	83
5	Yes	146	17	30	78
6	Male	1 780	18	0	95
7	Male	154		125	730
8	Female	141	170	21.43	87
9	Female	162	15.8	8	74
10	Male	210	18	32	58

2. Tom asked several students to help him measure the heights of a large group of Year 11 students. When he checked the data, he found that one height was recorded as 210 m.

   a. What would be the initial 'clean' Tom would apply to this data value?

   b. Even after he had 'cleaned' the data value, Tom noticed that this height was quite a bit taller than other heights recorded. Discuss what Tom could do as a result of this.

3. In a survey, some Year 13 students were asked how many hours of television they had watched in the previous week.

   a. Explain why a specific week was selected for this question, rather than asking the student to give an average number of hours they spend watching television each week.

   Two students replied zero to this question, for very different reasons.

   b. Suggest two very different reasons for a student watching zero hours of television in the previous week.

© ESA Publications (NZ) Ltd, Freephone 0800-372 266, ISBN 978-1-927297-63-6

Analysis

Various **statistics** can be calculated to summarise the features of a sample. For example, one way of describing an average value of a sample is to calculate the **mean** (sum of values divided by number of values).

The five summary statistics

The five key **summary statistics** used in the comparison of samples are:

- highest score (maximum)
- upper quartile
- median
- lower quartile
- lowest score (minimum).

Learn how to calculate these using technology.

Median

The **median** is the middle score (or the average of the two middle scores) when the data are **ranked** (put in order of size).

Example

The following are 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

Ranking the scores:

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

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

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.

Note: The **standard deviation** of a set of scores is another measure of spread that is often supplied when using software to produce statistics for data. The larger the standard deviation is, the more spread out values are about their mean.

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, 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).

So the five key summary statistics for the vocabulary test scores are

Minimum = 3

Lower quartile = 6

Median = 7

Upper quartile = 8

Maximum = 10

Note: The medians and quartiles divided data into quarters:

- $\frac{1}{4}$ (or 25%) of the data lies below the lower quartile
- $\frac{1}{2}$ (or 50%) of the data lies below the median
- $\frac{3}{4}$ (or 75%) of the data lies below the upper quartile

In the vocabulary test in the example above:

- the lowest 25% of scores were between 3 and 6
- the next 25% of scores were between 6 and 7
- the next 25% of scores were between 7 and 8
- and the top 25% of scores were between 8 and 10.

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. 46

Exercise C: Summary statistics

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. For the amount spent, find:

i. the lower quartile

ii. the upper quartile

c. 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

a. i. Complete the statement:

Half the sample scored ________ goals or fewer.

ii. What is the median for the sample?

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

i. the upper quartile

ii. the lower quartile

iii. 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 five key summary statistics for the following samples. Where necessary put data in order of size first.

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

Minimum

Lower quartile

Median

Upper quartile

Maximum

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

Minimum

Lower quartile

Median

Upper quartile

Maximum

Answers

Exercise A : Statistical variables and questions (page 2)

1. a. Gender (male, female)
   b. Age (years or nearest year)
   c. Country of birth (New Zealand, etc.)
   d. Ethnic group (Asian, Mäori, etc.)
   e. Height (centimetres)
   f. Right foot length (centimetres)
   g. Arm span (centimetres)
   h. Wrist circumference (centimetres)
   i. Popliteal length (centimetres)
   j. Index finger length (cm or mm)
   k. Preferred hand(s) (left, right, either)
   l. Transport method (bus, walk, car, etc.)
   m. Travel time (minutes)
   n. Weight of bag (kilograms or grams)
   o. Learning area/subject (Maths, PE, etc.)
   p. Reaction time (seconds)
   q. Sport/pastime (rugby, tennis, etc.)
   r. Fitness level (e.g. 1 (unfit) – 5 (very fit))
   s. Resting pulse rate (beats per minute)
   t. Length of time cell phone owned (months)
   u. Number of texts sent (whole number)

2. *Answers will vary – examples are given.*

   Comparisons could be made between the foot lengths of boys and girls of a certain age. For example, 'Do the foot lengths of 12-year-old boys tend to be longer than the foot lengths of 12-year-old girls in New Zealand?'

   Similar comparisons between boys and girls of various ages could also be made for the following variables:

   height, arm span; wrist circumference; popliteal length; index finger length; travel time to school; weight of bag; reaction time; resting pulse rate; length of time cell phone owned

   Alternatively, comparisons could be made between two different age groups, e.g. compare 10 year olds with 15 year olds (instead of boys and girls) for some appropriate variable.

3. a. Suitable: Height (cm or m) is being compared over male and female 10 year olds in New Zealand
   b. Suitable: Arm span (cm or m) is being compared over male and female Year 6 students in New Zealand.
   c. Suitable: Time taken to travel to school (minutes) is being compared for students who walk and students who travel by bus.
   d. Not suitable: Question involves 2 variables but only 1 group (**bivariate data**)
   e. Not suitable: Unless fitness can be given a numerical measure so comparisons can be made between Y9–11 students who walk to school and those who don't.
   f. Not suitable: Unless tidiness can be given a numerical measure which can be compared.
   g. Not suitable: One group and 2 variables (bivariate).
   h. Suitable: Resting pulse rate (beats per minute) is being compared for male and female Year 10 students in New Zealand.
   i. Not suitable: One group and 2 variables (bivariate).
   j. Suitable: Daily number of texts sent is being compared over two groups (Y11 boys and Y11 girls).

Exercise B: Cleaning data (page 6)

1. Cleaning of data:

   Row 2: time should be 75 (seconds)

   Row 3: height should be 170 (cm)

Row 4: delete (wrist measurement unlikely to be correct)

Row 5: delete (response ‘yes’ inappropriate for ‘sex’)

Row 6: height should be 178 (cm); query time to school (is student a boarder, for example?)

Row 7: wrist measurement missing and other entries inappropriate so delete this row of data.

Row 8: wrist measurement seems to be in mm (change to 17 cm); round time to school to 21 min

Row 9: round wrist measurement to 16 cm.

Row 10: check height (very tall, but possible).

2. a. Assume this is 210 cm (or 2.1 m)

   b. Tom could:

      i. remove this value as being incorrect, or

      ii. check with the student who did the measuring that the result was correctly measured and recorded, or

      ii. remeasure the student (who could be very tall).

3. a. The previous week is recent so students will usually remember more precisely than if asked to average out over an unspecified number of weeks – such averages can be very inaccurate.

   b. One student may have watched no television, or used any screen for entertainment during the week; another student may have watched many hours of television-type entertainment on their computer, tablet, etc.

Exercise C: Summary statistics (page 8)

1. a. $22 b. i. $18 ii. $27

   c. $9

2. a. i. 6 ii. 6.5

   b. i. 12 ii. 2 iii. 10

3. a. 9 b. 24 c. 29 d. 5

		Min	LQ	Median	UQ	Max
4.	a.	$23	$51	$92	$112.50	$130
	b.	$2.45	$3.65	$3.85	$4.20	$4.50
	c.	15	16	20	25	45
	d.	3.4 min	4.9 min	7.6 min	8.75 min	10.3 min
	e.	0.9 kg	3.15 kg	4.6 kg	7.4 kg	10.5 kg
	f.	29 cm	32 cm	33 cm	35 cm	50 cm
	g.	0	1	2	3	7

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

   e. Range = 9.6 kg IQR = 4.25 kg

   f. Range = 21 cm IQR = 3 cm

   g. Range = 7 IQR = 2

Exercise D: Dot plots (page 11)

1. a. There are two groups within the data: marks in the range 21–25 and marks in the range 29–34. The data in each of these groups is reasonably symmetrical within the group. There is an extreme value at 40 (someone got 100% in the test).

   b. There is a cluster of spending around $20 and another cluster of spending in the range $30–$35. Very few spent in central values $23–$29.

   c. There is one main group with two mounds. The iron levels within that group are distributed reasonably symmetrically about the value 17 µg/L. There are two outliers (above 36 µg/L); the iron levels of 28 µg/L and 31 µg/L are also higher than most.

   d. The data is fairly uniform (similar numbers of reaction times from 0.3 sec to 1.3 sec). Most of the reaction times were between 0.5 sec and 1 sec, with two faster and four slower than this.

   e. The distribution of egg weights is approximately symmetrical and bell-shaped. Most egg weights are clustered between 25 g and 30 g, with a peak at 27 g (the mode). There are a couple of lighter eggs (20 g and 22 g) and a couple of heavier eggs (32 g and 35 g).

INDEX

analysis (statistical) 2

back-to-back stem-and-leaf plot 16
bar (or column) graph 13
bimodal (distribution) 10
bivariate data 45
box 18
box-and-whisker plot 18

census 24
cleaning data 5
clusters 10
comparison (statistical) 1
composite bar graph 13
conclusion (statistical) 36

data (statistical) 2, 5
database 1
'dirty'data 5
distribution 10, 18
dot plot 10

gaps (data distribution) 10

histogram 13

inference 31
interquartile range 7

key 16

leaf 16
lower quartile (LQ) 7

mean 7
median 7
multivariate data set 1

negatively skewed 20

ordered stem-and-leaf plot 16
outliers 10
overall visual spread 31
overlap 31

pie graph 13
population 1, 24, 31
positively skewed 20
PPDAC 1, 36

random sample 1, 24
range 7
ranked (data) 7

sample 5, 24
sampling variability 24
shift 31
skewed (left/right) 10, 20
standard deviation 7
statistical enquiry cycle (PPDAC) 1, 36
statistical questions 2
statistics 7
stem 16
stem-and-leaf plot 16
summary statistics 18
symmetrical (distribution) 10, 20

tables (of data) 5
terms (statistical) 2

uniform (rectangular) distribution 10
unimodal distribution 10
upper quartile (UQ) 7

variable (statistical) 1, 2

whiskers (box-and-whisker plot) 18

© ESA Publications (NZ) Ltd, Freephone 0800-372 266, ISBN 978-1-927297-63-6