7 Plan the Investigation 3

A Describing Variables and Managing Sources of Variation

As part of your plan for your investigation you will have sourced data and identified the variables of interest. You should also state the source of your data in your investigative report and consider the sources of variation that exist as part of the collection of the data. You may not have collected the data yourself and so you will need to consider the possible sources of variation that may have occurred in the original collection process and how they may have been managed.

Examples of possible sources of variation could be : the weather, the amount of assistance given, the instructions provided, the equipment that is used, and the ability of the person taking the measurements.

Sources of Variation Categories

Sources of variation can be considered in five different categories :

1 **Natural variation** : this is the differences that occur due to nature, such as the variation in heights or head circumference between different people.

2 **Occasion to occasion variation** : this type of variation occurs when repeated measurements are taken over time. For example, a person's blood sugar levels will not be the same at different times of day.

3 **Measurement variation** : when taking a measurement, variation can occur in the method. For example, when measuring a height, it would be important to make sure the person being measured removes their shoes as this could add a different amount to each height due to the different sized soles and heels on shoes worn by different people.

4 **Induced variation** : different factors may affect the same measurement. For example, the price of fuel in two different petrol stations in different areas of the country may be affected by their location relative to the nearest port. They could also be part of different companies. Care needs to be taken when considering this type of variation in the context of comparison investigations.

5 **Sampling variation** : this always occurs when a sample is taken. Unless the entire population of interest is measured, sampling variation will be present. This means that each time a different sample is taken, variation will occur in the individuals present in the sample and therefore the sample statistics will vary between samples.

Example :

For the investigative question : *'I wonder if the handspan of the students who play netball in Girls' High School tends to be greater than the handspan of those who don't play netball in Girls' High School'.*

a) **Describe** the variables and **how** they will be measured.

b) **Describe** possible sources of variation in the measurements and what actions that would have to be taken to ensure the data is consistent.

c) Where possible **describe** what would happen if these actions were not taken.

Possible Answers :

a) *The two variables are the student's hand span and whether they play netball or not. Hand spans (in cm) will be measured by getting people to fully stretch their hand along a ruler and measuring from the outside of the little finger to the outside of the thumb. Whether they play netball or not will be determined by asking them if they have played at least 5 games for a team in the last year. The answer will be either "Yes" or "No".*

b) *When measuring hand spans people need to stretch their hand as much as possible. Measurements need to be taken from the same part from the finger to the same part of the thumb each time. This is to minimise measurement variation.*
- Natural variation will be present as handspans naturally vary between people.
- Students will be asked which team they played for and position. This is also to minimise measurement variation.
- Sampling variation will also be present if a sample of students is taken from the population of students at Girls' High School as part of this investigation.

c) *If students do not stretch their hand out fully then that could result in the measurements taken being less than they should be. Inconsistency in the position of the finger or thumb measured from could lead to measurements being under or over what they should be. Students could misremember when they last played netball causing them to be grouped wrongly. Asking them to say what team they played for and position will hopefully help them be more accurate in their recollections.*

1 For each of these investigations state how the variables would have to be measured. Discuss possible sources of variation and actions that could be taken to ensure the data is consistent.

a) *'I wonder if the time taken to run 100m tends to be faster for Year 13 students in Girls' High School than the time taken to run100m for Year 12 students in Girls' High School?'*

i) Describe the variables and how they will be measured.

……

……

……

……

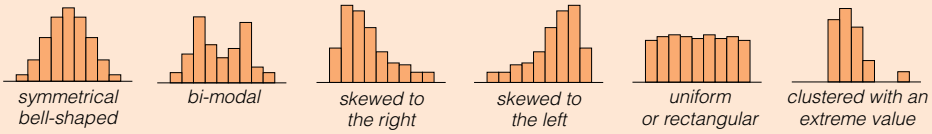
13 Data Display 3

A Describing a Distribution

The shape of a histogram reveals the distribution of the data.

Interesting details of the distribution could be the peak score (mode), gaps, extreme values, or clusters.

Vocabulary to describe the shape of distributions :



1 The histogram shows weights of checked in luggage on an Air New Zealand flight.

a) Name and describe the shape of the distribution. .....

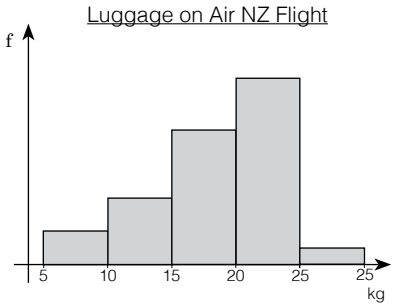
.....

b) Can you think of a reason for the data to have this shape? .....

.....

.....

.....



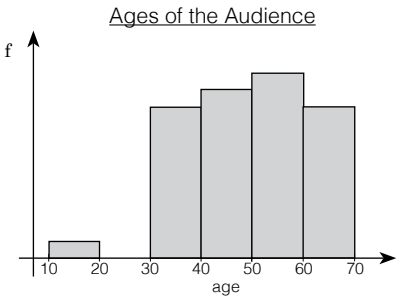
2 Ages of the audience of Shakespeare's play *'Hamlet'* are displayed in a histogram. Describe the shape and interesting features of the distribution.

.....

.....

.....

.....



3 Students of Year 11 at Bay View College drew the four graphs i) to iv) below. The four graphs have no titles, no labels, no scales. In the grey boxes are the titles, together with descriptions of the graph.

a) For each graph select the correct title - A, B, C or D.

b) Place labels and scales on the horizontal axes.

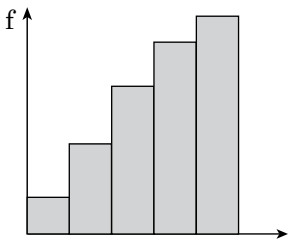
A **Travel Time from Home to School**
Times are clustered between 0 and 30 minutes with an extreme time of 40-50 minutes.

B **Number of Students per Class**
A triangular distribution, with class sizes ranging from minimum 5 students to maximum 30 students.

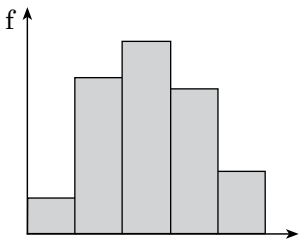
C **Time Students Spend in the Library at Lunchtime**
A bimodal distribution with most common times 0-5 mins and 15-20 minutes.

D **Heights of Year 11 Boys**
Almost symmetrical bell shaped distribution centred at 170-180 cm, ranging from 150 cm to 200 cm.

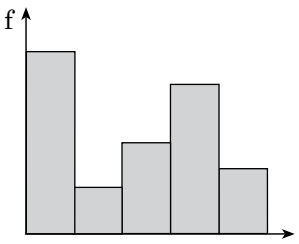
Graph i) ; Title



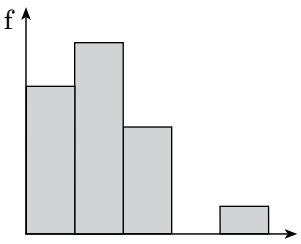
Graph ii) ; Title



Graph iii) ; Title



Graph iv) ; Title



21 Measures of Centre and Spread 4

A Measures of Spread

Apart from a measure indicating the centre we also like a measure indicating the spread of a given data set. **Range** and **interquartile range** (**IQR**) are measures of spread.

Range = maximum score minus minimum score;
the range tells us how widely spread the scores are overall.

Interquartile range = upper quartile minus lower quartile;
the interquartile range tells us the spread of the middle 50% of the scores.

In short : **Range = Max – Min** **IQR = UQ – LQ.**

Example : Here is an ordered set of 17 scores.

12, 14, 16, 17, 18, 19, 20, 20, 21, 22, 22, 22, 23, 24, 24, 25, 28

a) Find median and quartiles.
b) Calculate range and interquartile range.
c) Draw a dotplot; on it indicate median and interquartile range.

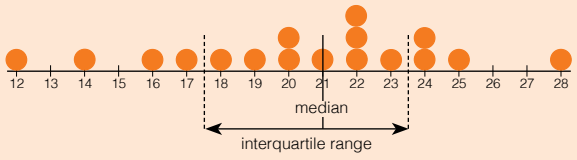
Working :

12, 14, 16, 17 | 18, 19, 20, 20, 21, 22, 22, 22, 23 | 24, 24, 25, 28

a) Med = 21, LQ = 17.5, UQ = 23.5

b) Range = 28 – 12 = 16 IQR = 23.5 – 17.5 = 6

c)



1 Members of an orchestra were asked at what age they had started playing their instrument. This is the data :

4	6	7	7	8	8	8	9	9	10	10	10
10	11	11	12	12	12	14	14	14	15	17	22

a) Work out median and quartiles.

……

……

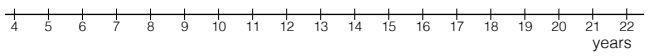
b) Calculate range and interquartile range.

……

……

c) Make a dot plot. On the bottom show the median and interquartile range.

Musicians Age when Starting to Play



B The Five Summary Statistics

Minimum and maximum value, median, lower and upper quartile are called **'the five summary statistics'** of a sample.

We can use a graphic calculator to work out these values.

Here are the instructions for the ***Casio fx 9750 or 9860 GIII*** :

Enter via **STAT** on the main menu.

Before starting, delete any old data from list 1 with **DEL-A** (F6 then F4)

We enter single scores into list 1. Therefore in the **CALC** menu we must **SET** the

1Var XList to List1 and 1Var Freq to 1 (press **EXE**).

Enter the scores into List 1, then simply select **CALC** then **1Var** to get all the statistical measures needed.–

In the displayed list you find the mean ($\bar{x}$), and when you scroll down you find minimum value, lower quartile, median, upper quartile and maximum value (**minX**, **Q1**, **Med**, **Q3**, **maxX**).

Example :

Enter the data of column A, question 1. After pressing **CALC**, scan the list and find $n = 24$ (n gives the number of data values, always use this to check the data entry), scroll down to find :

minX = 4, **Q1** = 8, **Med** = 10, **Q3** = 13, **maxX** = 22.

1 These are heights (in cm) of a randomly selected sample of fashion models.

172	168	171	175	167	162	167	180	166
170	169	169	174	172	175	177	178	170
165	172	173	173	160	181	178	173	167
175	169	164	174	171	178	175	165	168

a) Enter the data into your GC. Check the value of n.

n = …… , is that correct? ……..

b) Find the five summary statistics in centimetres.

min = ……., LQ = ……., Med = ……,

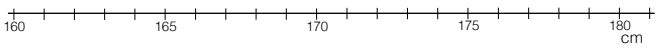
UQ = ……., max = ……..

c) Calculate range and interquartile range.

Range = …… ; IQR = ……

d) Make a dot plot of the heights of fashion models and show the median and interquartile range.

Heights of Fashion Models



27 Box-and-Whisker Plots 1

A Drawing Box and Whisker Plots

To show the **frequency** of the scores in a distribution we can draw a **dot plot**, a **bar graph** or a **histogram**.

To show the **spread** (or **variation**) of the distribution we draw a **box-and-whisker plot** and to do this we need **the five summary statistics**.

Example : Listed on the right are the weights of players in a senior regional rugby squad.

a) Draw a dot plot for the data.

b) Find the five summary statistics.

c) Draw a box-and-whisker plot.

Weights of the Senior Rugby Squad (kg)

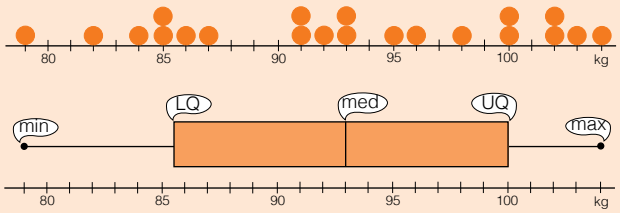
79	82	84	85	85	86	87	91
91	92	93	93	95	96	98	100
100	102	102	103	104			

Answer : a) See top plot.

b) min = 79, LQ = 85.5, Med = 93, UQ = 100, max = 104

c) See bottom plot.

Weights of the Senior Rugby Squad



Things to note :

The box-and-whisker plot has 4 sections. Each section contains the same number of scores. When the scores are close together the section is short, when the scores are spread out the section is long.

1 The fruit and veges buyer of our local supermarket wrote down the number of pineapples sold each day in July.

This is his list :

16	20	10	18	23	24	26	22	29	30	21	18	14
20	24	26	23	25	25	34	37	28	23	20	24	17
15	12	14	16	10								

a) Draw a dot plot for the data.

b) Give the five summary statistics :

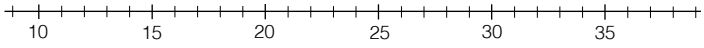
min = ……, LQ = ……, Med = ……

UQ = ……, max = ……

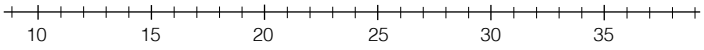
c) Draw a box-and-whisker plot for the data.

Pineapples Sold per Day in July

(dot plot)



(box-and-whisker)



2 Ages of a sample of 25 bowling club members are displayed in the stem-and-leaf plot.

Draw a box-and-whisker plot to show the spread of the data.

……

……

……

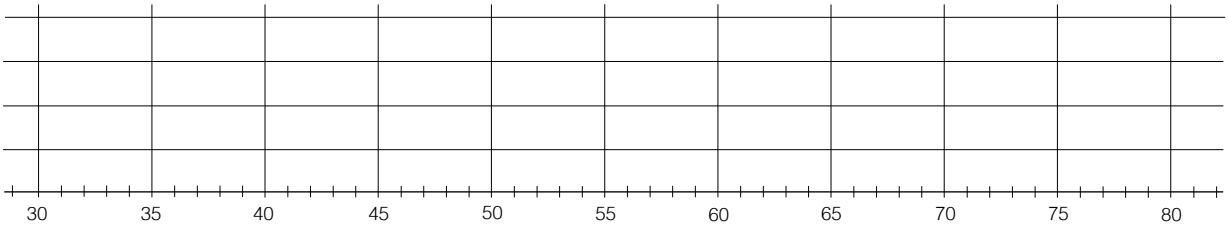
……

Ages of Bowling Club Members

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

key 3 \| 5 = 35 yrs

Ages of Bowling Club Members



37 Statistical Inference 4

A Making the Call - Overall Visible Spread

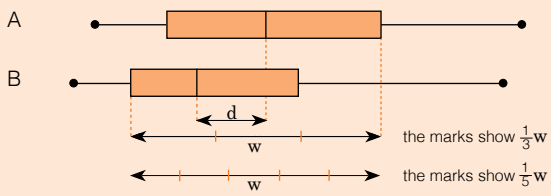
If the Three Quarters/Half Rule fails, use the Overall Visible Spread Rule. These are the steps for using the Overall Visible Spread Rule :

Steps : i) Look at distance between the two medians (=$\mathbf{d}$).

ii) Look at the overall visible spread, that is the total width of the boxes together (=$\mathbf{w}$).

Rule : With a sample size of 30, you can make the call that the scores of population A tend to be larger on average than those of population B, only if $\mathbf{d} > \frac{1}{3}$ of $\mathbf{w}$.

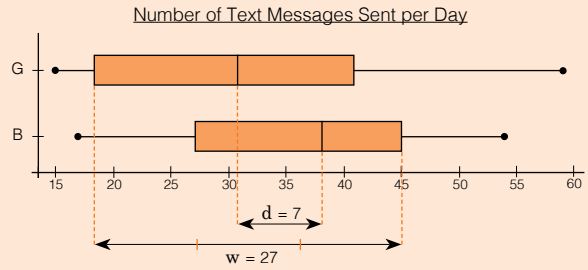
With a sample size of 100, the call can be made if $\mathbf{d} > \frac{1}{5}$ of $\mathbf{w}$.



Example :

Gemma and Benjamin recorded how many text messages they sent each day in the previous month ($\mathbf{n}$ = 30). The box-and-whisker plot show the results.

Is there sufficient evidence to conclude that in general Benjamin tends to send more text messages per day than Gemma?



Conclusion :

The difference between the medians ($\mathbf{d}$) is about 7 messages, the overall visible spread ($\mathbf{w}$) is about 27 messages.

Since $\frac{1}{3}$ of 27 = 9, $\mathbf{d}$ is not larger than $\frac{1}{3}$ of $\mathbf{w}$, the difference between the medians is not large enough to make the call. The shift shown here could just be due to sampling variation. If we would take repeated samples of 30 days, each would give a different picture and Benjamin's number of text messages may not always be further up the scale than Gemma's. Back in the population of all daily messages sent by Gemma and Benjamin, the pattern shown here is not necessarily happening.

There is *insufficient evidence* to conclude that back in the populations, Benjamin tends to send more text messages per day than Gemma.

1a) Look back at the example. What if Gemma and Benjamin had recorded their scores for 100 days in stead of 30, and the above box-and-whisker graph showed their results over that time. Would the conclusion be different? Show your working.

.....

.....

.....

.....

b) What is the reason that there are different criteria for different sample sizes?

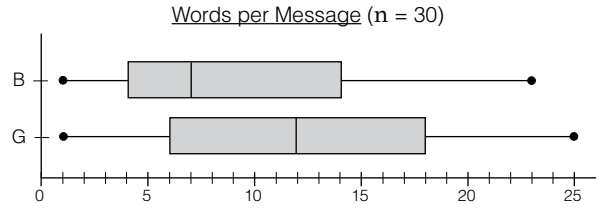
.....

.....

.....

.....

2 Gemma and Benjamin did a survey to see whether Gemma sends longer text messages (more 'words') than Benjamin. They took random samples of 30 text messages on their phones. Find $\mathbf{d}$ and $\mathbf{w}$ and write a conclusion about who sends longer texts.



.....

.....

.....

.....

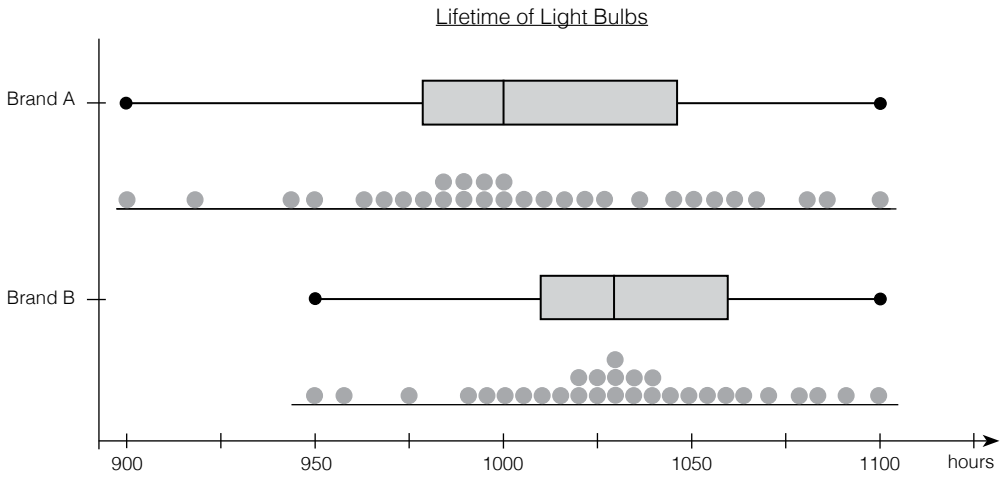
.....

.....

41 Form a Conclusion 2

A Bright Lights

1 A consumer organisation is investigating the lifetimes of two brands of light bulbs. Random samples of 30 bulbs of each brand were taken and the time until each bulb blew was recorded. The plots below show the results.



a) Which brand of light bulb seems to have more predictable lifetimes. Give a reason for your answer.

.....

.....

b) Discuss overlap and shift of lifetimes in the samples.

.....

.....

.....

.....

.....

.....

c) What shape do you expect a histogram of the population of *'lifetimes of light bulbs'* to have? Explain.

.....

.....

.....

.....

2 Can we make a claim when comparing the lifetimes of these light bulbs. Explain.

.....

.....

.....

.....

.....

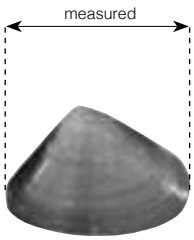
.....

.....

45 Practice Investigation 1

A Shell Sizes

1 Tim has two favourite places where he gathers pipis. He reckons site A has larger pipis than site B. Tim investigates his assertion by collecting 30 pipis from each site. He measured them to the nearest millimetre, at the widest part of the shell.



Pipi Shell Sizes (mm)

Site A				
57	64	65	71	54
46	74	58	53	35
48	62	76	57	58
72	43	44	60	38
62	59	41	57	47
67	38	44	55	54

Site B				
53	56	49	68	60
62	57	47	65	47
34	59	63	51	43
40	37	54	59	58
53	45	39	47	56
41	51	43	58	44

Write an investigative question for Tim.

.....

.....

.....

2 Comment on the possible sources of variation present in the collection of this data.

.....

.....

.....

3 Draw a back-to-back stem-and-leaf plot for the data. (Use scrap paper first, then order the leaves and copy the result into the space provided.)

Shell Sizes

Site A	Site B

key :

4 Calculate summary statistics.

	Site A	Site B
mean		
minimum		
LQ		
median		
UQ		
maximum		
IQR		

Page 3 - Define the Problem 1

A1 Possible answer, many others are possible.

a) I wonder, do NZ athletes competing in rowing and canoeing tend to be heavier than NZ athletes competing in running?
Populations : (A) Olympic athletes competing in rowing and canoeing;
(B) Olympic athletes competing in running
Variable : weight of athlete

b) I wonder, in NZ, do female Olympic athletes tend to be younger than male Olympic athletes?
Populations : (A) female Olympic athletes (B) male Olympic athletes
Variable : age of the athlete

Page 4 - Define the Problem 2

A1 Possible answer, many others are possible.

a) I wonder, do male drivers caught speeding between Christchurch and Ashburton tend to drive faster than female drivers caught speeding?
Populations : (A) male drivers caught speeding between C and A; (B) female drivers caught speeding between C and A. Variable : speed (km/h)

b) I wonder, do drivers caught speeding between Christchurch and Ashburton who are younger than 34 tend to go faster than speeding drivers 35 or over?
Populations : (A) drivers younger than 35 caught speeding between C and A; (B) drivers aged 35 or over caught speeding between C and A.
Variable : speed (km/h)

c) I wonder, do speeding drivers who drive alone between Christchurch and Ashburton tend to go faster than speeding drivers who have passenger(s) in the car?
Populations : (A) drivers driving alone caught speeding between C and A; (B) drivers with passenger(s) caught speeding between C and A.
Variable : speed (km/h)

d) I wonder, do speeding drivers caught in the afternoon between Christchurch and Ashburton tend to go faster than speeding drivers who are caught in the morning?
Populations : (A) drivers driving alone caught speeding between C and A; (B) drivers with passenger(s) caught speeding between C and A.
Variable : speed (km/h)

Page 6 - Plan the Investigation 2

A1 No. The samples are large enough (about 30), but it is unlikely that these students represent all 12 year olds in New Zealand. Height could be linked to ethnicity and it is unlikely that the students of one school fairly represent the ethnic make up of New Zealand's 12 year olds.

A2 No. Possible comments (other are possible) :
i) Blake's two samples are selected on two different days of the week. There may be different patterns in commuting habits on different days.
ii) Blake's sample only has commuters who end up in the CBD; commuters who get off on smaller stations are not included in the sample.

A3 a) Dairy farmers need to tell him how many milk producing cows there were that day.

b) He could select three days in every month, say the 1st, 10th and 20th of every month, on which he records the milk production plus number of cows. This gives him 36 scores from each farm. If for some reason he misses out a few, he would still have at least 30. Alternatively, he could daily record this data for one entire month; this would take less time but may not be a fair comparison.

A4 Lucy has access to cats that go to her mother's vet clinic. Sick cats would be over-represented and sick cats could be either very thin or maybe obese.

Pages 7-8 - Plan the Investigation 3

A1 a) i) The variables are: the time taken to run 100m and the year group of the student. The time will be most likely measured with a stopwatch and the year group determined by the school.

ii) Variation in times could be caused by the weather, the terrain run over, the accuracy of the timings, how hard the runners try etc. This could be managed by taking times from a school sports day for example, where the race track, weather and measuring of time is likely to be consistent. The year groups should be reasonably consistent.

b) i) Reaction time is measure in milliseconds. Drivers license will be yes or no for those having either a restricted or a full license (or could decided to only consider students will full licenses as having a license. Easiest way to measure reaction time is using an online site designed to measure this. License or not will be determined by asking the student.

ii) Variation could occur if students taking the reaction test use different computers, are distracted by other students, have different numbers of turns (so get more practice). This can be solved by doing the test in a quiet environment with the same computer for each student and taking the average of the first three turns.

c) i) A blade of grass will be measured in mm using a ruler. The blades of grass for the sample will be cut using a pair of scissors and then measured. The front/back of the school areas will be determined using a map.

ii) Variations could occur in the way the grass is cut and measured. This should be done using the same sharp pair of scissors, in line with the ground. Use the same person measuring each time to ensure consistency and a ruler with clear marks for each mm.

Page 9 - Plan the Investigation 4

A1 Faulty data :

i) cell C4 - typo; change to 'newzealand'
ii) cell J4 - 0 minutes seems unlikely, even if you live next door; flag with X
iii) cell G7 - measurement given in cm, could change to 80 or X
iv) cell H7 - same problem, change to 70 or X
v) cell K8 - unlikely that a 15 year old girl from the Philippines carries a 15000g = 15 kg bag, change to 1500
vi) cell M8 - change to X
vii) cell A10 - typo; change to female
viii) cell E10 - measurement given in metres, change to 167
ix) cell I11 - left blank, change to X
x) cell F14 - impossible figure, most likely this should be 25
xi) cell M15 - I think this boy doubled his count, change to 72
xii) cell D16 - misunderstanding, change to 1

Pages 11-12 - Data Display 2

A1

Heights of Athletes

Water, Road & Field Sports		Ball Sports
7	15	
9 8 8 8 6 4 3 0	16	4 4 5 5 6 7
8 8 6 6 6 5 3 1 0	17	0 2 3 4 4 5 7 9
7 5 3 3 3 3 2 1 0 0	18	0 2 2 3 3 5 5 6 7 7 7
9 9 5	19	0 5 6 6
	20	1

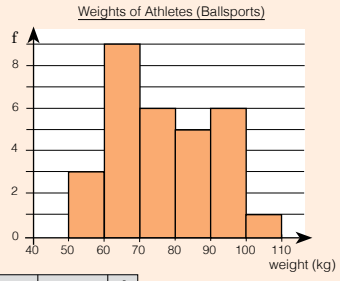
key : 19|5 = 195 cm

A2 a)

weight	tally	f
40 - (50)		
50 - (60)	///	3
60 - (70)	𝍸 ////	9
70 - (80)	𝍸 /	6
80 - (90)	𝍸	5
90 - (100)	𝍸 /	6
100 - (110)	/	1

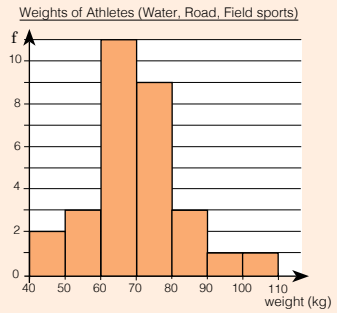
Pages 11-12 - Data Display 2 - continued

A2 a)



b)

weight	tally	f
40 - (50)	//	2
50 - (60)	///	3
60 - (70)	𝍸 𝍸 /	11
70 - (80)	𝍸 ////	9
80 - (90)	///	3
90 - (100)	/	1
100 - (110)	/	1



B1 a) Ball Sports

age-group	tally	f
15 - (20)	//	2
20 - (25)	𝍸	5
25 - (30)	𝍸 𝍸 𝍸 /	16
30 - (35)	𝍸	5
35 - (40)	//	2
40 - (45)		

Water, Road, Field Sports

age-group	tally	f
15 - (20)	///	3
20 - (25)	𝍸 ////	9
25 - (30)	𝍸 𝍸 /	11
30 - (35)	𝍸	5
35 - (40)	/	1
40 - (45)	/	1

b)