Reviewing Statistics

Throughout your study of strand 1 you will have considered all aspects of a statistical approach:

- asking a question that results in data that varies
- displaying this data in a way that allows you to see patterns in the variation
- analysing the patterns in the data
- drawing conclusions from that data.

You may even have had an opportunity to get a glimpse of what it is like to become a statistical detective; attempting to account for unexpected variability you observe in a particular set of data.

As you review for the final examination in June, it is important that you can connect each element of your study and consider the BIG IDEA of the strand so that you will be able to use the elements appropriately to help you solve problems that you may not have seen before.

The following is an extract from Strand 1 of the syllabus; it summarises what you should be able to do when you finish studying this strand.

> It is envisaged that throughout the statistics course students will be involved in identifying problems that can be explored by the use of appropriate data, designing investigations, collecting data, exploring and using patterns and relationships in data, solving problems, and communicating findings. This strand also involves interpreting statistical information, evaluating data-based arguments, and dealing with uncertainty and variation.

You may decide to form a study group with your friends or you may prefer to work alone; either way as you work through this review document you will consider issues such as framing a question in order to obtain meaningful ***reliable*** data, selecting a sample in order to avoid ***bias, displaying*** your data in a way that will allow you to see patterns in the variation and ***drawing conclusions*** from your data.

Asking the Question

Think 

Do you use a computer?

How did you answer the question?
What were you thinking when you answered it?

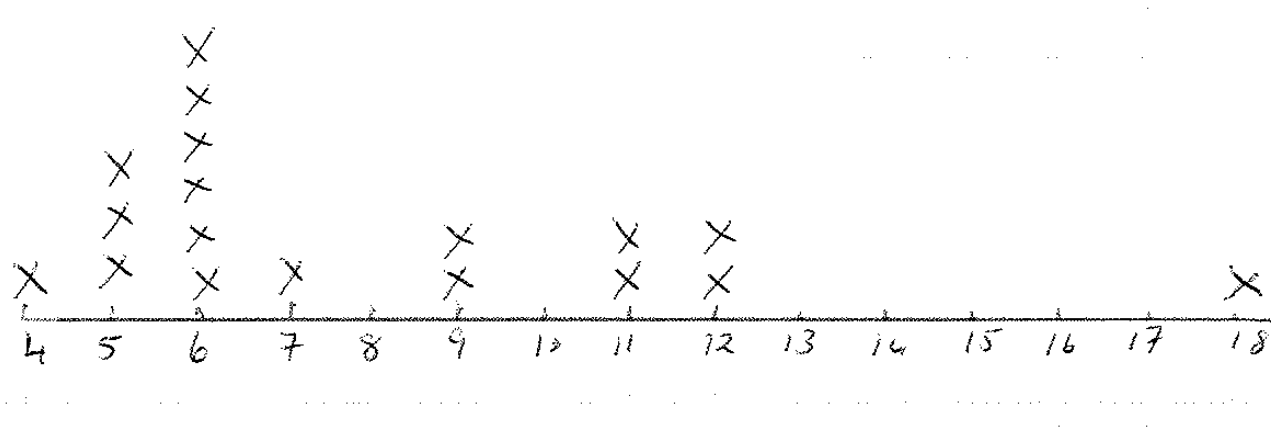
A university sports outlet was considering shutting down their campus shop and becoming an on-line store in an effort to reduce costs. A group of students was surveyed and asked that same question:

Do you use a computer?

Sophie answered **Yes** because she thought the question meant had she ever used a computer.
Joe answered **No** because he thought the question was asking whether he used one regularly.
Andrew answered **No** because he played games on the computer and didn't think this counted as "using" one.
Do you think the results of this survey are ***reliable?***

How could you rephrase the question so that it is less ambiguous and more likely to provide useful answers?

A group of students interested in finding the typical family size for their class obtained the data displayed in this line plot



What question do you think they asked in order to elicit this data?

What issues would they have needed to consider when framing the question?

Displaying the data and drawing conclusions from it

Use fractions or percentages to describe the data.

Can you see any clumps or areas where a large proportion of the data falls?

Are there any unusual family sizes? [18 is an unusual value in this set.]

What do you think is the typical family size of this group ? Why?

If you were asked to predict the family size of someone from this group what value would you give? Why?

How certain would you be? Can you lower this to a smaller range? How **confident** would you be now?

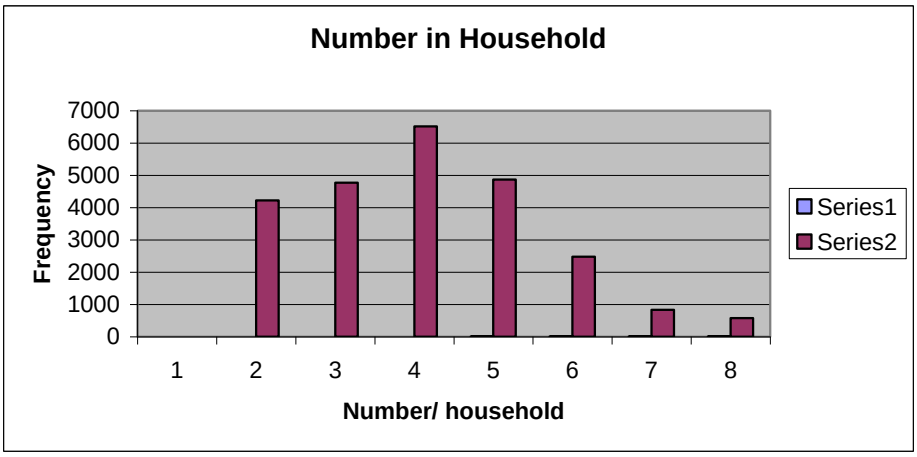
Calculate the mean family size for this group and identify the median family size. Which is a more reasonable estimate of typicality?

You could do a similar survey of your class, display the data in a line plot and compare the two data sets.

Or you could visit

http://beyond2020.cso.ie/Census/TableViewer/tableView.aspx?ReportId=109241 and retrieve some data from the area in which you live, use Excel to display the data and compare it to the sample above.

Household size	Frequency
2	4218
3	4773
4	6512
5	4870
6	2478
7	833
8	572

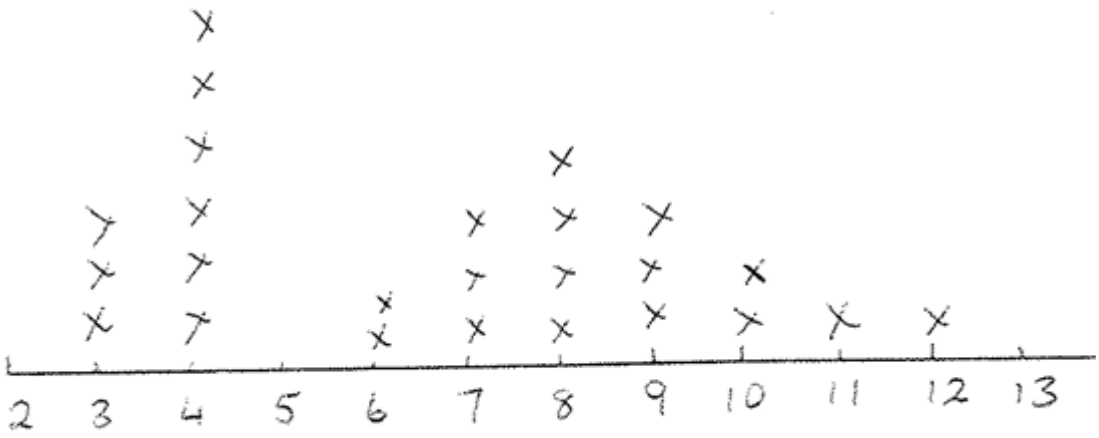


This bar chart was drawn with data from Carlow.

Compare this data with that from the sample set above. What is the range of this data set? What is the range from the sample data set?

Is there any evidence to suggest that the sample was from Carlow? Explain.

The following data set was gathered from a TY class who were interested in finding out what was the typical amount of money spent by their parents on the lotto each week.



Use fractions or percentages to describe the data.

Can you see any clumps or areas where a large proportion of the data falls?

Are there any unusual amounts?

What do you think is the typical amount spent on the Lotto each week by this group ? Why?

If you were asked to predict the amount spent on the Lotto each week by someone from this group what value would you give? Why?

How certain would you be? Can you lower this to a smaller range? How **confident** would you be now?

Return to the value you think is the typical amount spent on the Lotto each week by this group.

Calculate the mean amount spent on the Lotto by this group and identify the median amount spent on the Lotto each week. Which is a more reasonable estimate of typicality?

The data below was obtained by students trying to find out how good they are at judging a minute.

18	25	26	30	40	41
45	45	47	52	52	56
67	68	74	79	109	

Think 

Will a line plot be a meaningful display for this data? Try it out and see?
Now display the data in a **stem and leaf plot.** Which display is more meaningful? Why?

What is the range of the data? Where is all the data concentrated? Calculate the percentage of data in this region. How good were people at guessing a minute?
Predict how long do you think people in this group think a minute is? How **confident** are you of this answer?

Now consider **Q4** on the **2010 FL** mock paper, which requires you to use those skills you've been practicing.

Question 4 **(25 marks)**

Sam wanted to see how well the people in his class could judge how long one minute is.
He asked each student to say *'Start'* and then to say *'Stop'* when they thought a minute had gone by.
Sam used a stop-watch to time each of them.
He recorded the results in the table below.

Times in seconds					
63	56	86	52	75	65
57	59	64	55	89	54
39	67	82	70	68	57
66	72	33	42	52	79
60	59				

(a) Display the data in a stem and leaf plot.

(b) How many people did Sam time?

(c) What is the median time?

Extension to OL

Compare Sam's set of data with the previous set.

- Use two **stem and leaf plots** or **back-to-back stem and leaf plots.**
- Use % and fractions.
- Mention areas where the data is clumped.
- State the range of both sets.

Which group do you think was better at estimating a minute?

Justify your answer.

Q9A a) and **b)** on the **2010 OL** mock paper

Oxygen levels in a polluted river were measured at randomly selected locations before and after a clean-up. These results are given in the table.

Before (mg/l)				After (mg/l)			
20	25	20	9	26	10	10	9
23	23	10	11	11	15	11	11
2	10	11	5	3	8	11	4
11				13			

(a) Construct a back-to-back stem-and-leaf plot of the above data.

(b) State **one difference** and **one similarity** between the distributions of the measurements before and after cleanup.

Difference:

Similarity:

Q7 a) and **b)** on the **2010 HL** mock paper

Some research was carried out into the participation of girls and boys in sport. The researchers selected a simple random sample of fifty male and fifty female teenagers enrolled in GAA clubs in the greater Cork area. They asked the teenagers the question: *How many sports do you play?*

The data collected were as follows:

Boys	Girls
0, 4, 5, 1, 4, 1, 3, 3, 3, 1,	3, 3, 3, 1, 1, 3, 3, 1, 3, 3,
1, 2, 2, 2, 5, 3, 3, 4, 1, 2,	2, 2, 4, 4, 4, 5, 5, 2, 2, 3,
2, 2, 2, 3, 3, 3, 4, 5, 1, 1,	3, 3, 4, 1, 6, 2, 3, 3, 3, 4,
1, 1, 1, 2, 2, 2, 2, 2, 3, 3,	4, 5, 3, 4, 3, 3, 3, 4, 4, 3,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3	1, 1, 3, 2, 1, 3, 1, 3, 1, 3

(a) Display the data in a way that gives a picture of each distribution.

(b) State **one difference** and **one similarity** between the distributions of the two samples.

Difference: _____

Similarity: _____

(c) Do you think that there is evidence that there are differences between the two populations? Explain your answer.

Note: you are not required to conduct a formal hypothesis test.

Answer: _____

Justification:

Q 7 OL 2010 mock paper

(a) A teacher asked the students in her class to estimate the height of the church opposite the school in metres.

The stem-and-leaf diagram shows all the results:

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

Key: 3 | 5 represents 35 m

(i) How many students were in the class?

(ii) Describe the **shape** of the distribution of the data.

(iii) What was the median estimate?

(iv) Explain the answer to part **(iii)** to someone who does not know what the word "median" means.

Stop and think

Under what conditions would a **line plot** be a meaningful representation?

Under what conditions would a **stem and leaf plot** be a more meaningful representation?

Try using statistics to solve this problem.

PROBLEM: Climbing helmets are made in a variety of styles and sizes.
The manager of You Climb Safely must decide what styles of helmet to keep in stock and how many helmets of each size to order. A standard fit helmet is offered in 10 sizes. When you order helmets you must order 1000. How many of each helmet size should the manager order?

In order to get an idea of how head sizes are **distributed** the manager decided to measure the head circumferences of a group of people.

Think: what is the **population** of interest? Can he measure the circumferences of the heads of the whole population? How will he choose a **sample**?

The manager chose a **Simple Random Sample** of climbers from clubs around the country and recorded their head circumference and gender ion the table overleaf.

Is this a suitable sample? Why or Why not?

Gender	Head Circumference (mm)
F	522
M	580
M	552
F	531
M	563
F	546
F	545
M	545
M	545
M	568
F	560
M	613
F	555
F	573
M	585
F	584
M	600
M	595
M	593
F	590
M	594
F	564
F	536
M	586
F	540
M	585
M	550
M	565
F	600
F	590
F	551
M	590
M	580
F	577

Is a line plot a good representation of this data?

Display the data in a stem and leaf plot.

Describe the data…..Are there any clumps or areas where the data is concentrated? Are some head sizes more common than others?

Use your representation to answer the original question: ***how many helmets of each size should the manager order?***

Begin by counting the number of leaves on each stem.

Look at the first stem…52 ..How many leaves are there on stem 52? What fraction of the total is this? What % of the total number of head circumference measurements does stem 52 represent?

How many helmets size 520cm- 530cm should the manager order?

Continue working like this until you have decided how many helmets of each size the manager should order.

Return to your representation…Do you think there are **gender effects?** Try representing the male and female data in ***back to back stem plots*** Compare the two sets of data; is there any evidence to suggest that there are differences in the sizes of heads of men and women?

If there are gender effects will this affect the number of helmets the manager should order? Or are helmets unisex?

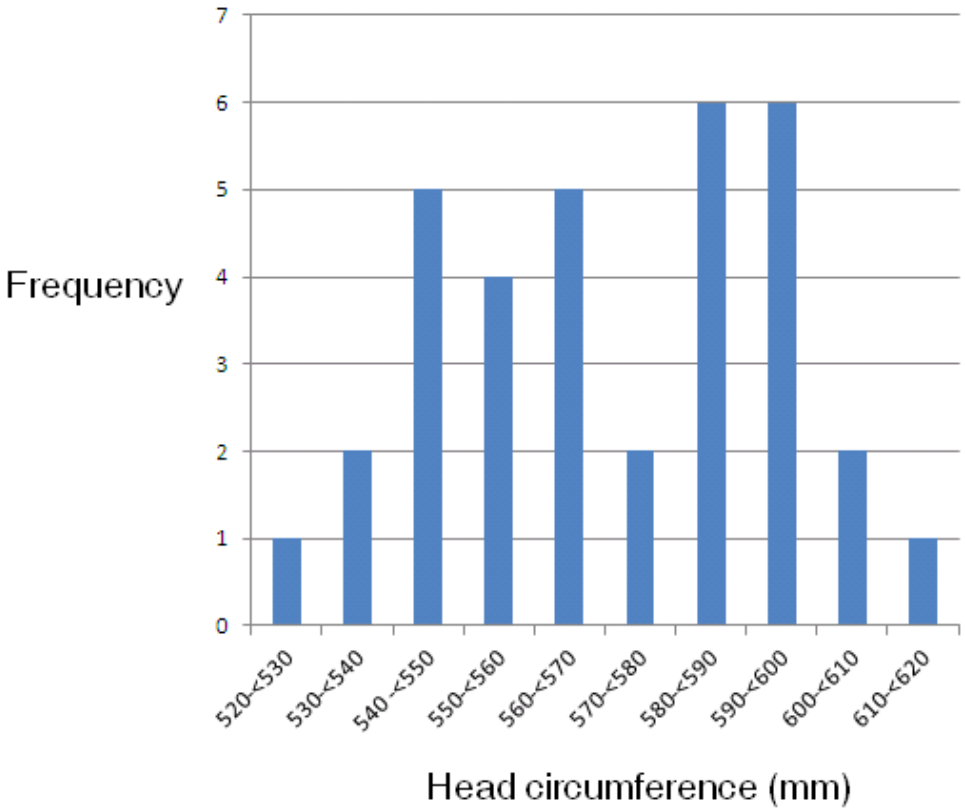
So far you have looked at **line plots** and **stem and leaf plots.** Both are very useful representations for allowing you to see patterns in the variation of your data. A histogram is another useful representation and it is especially useful when dealing with lots of data.

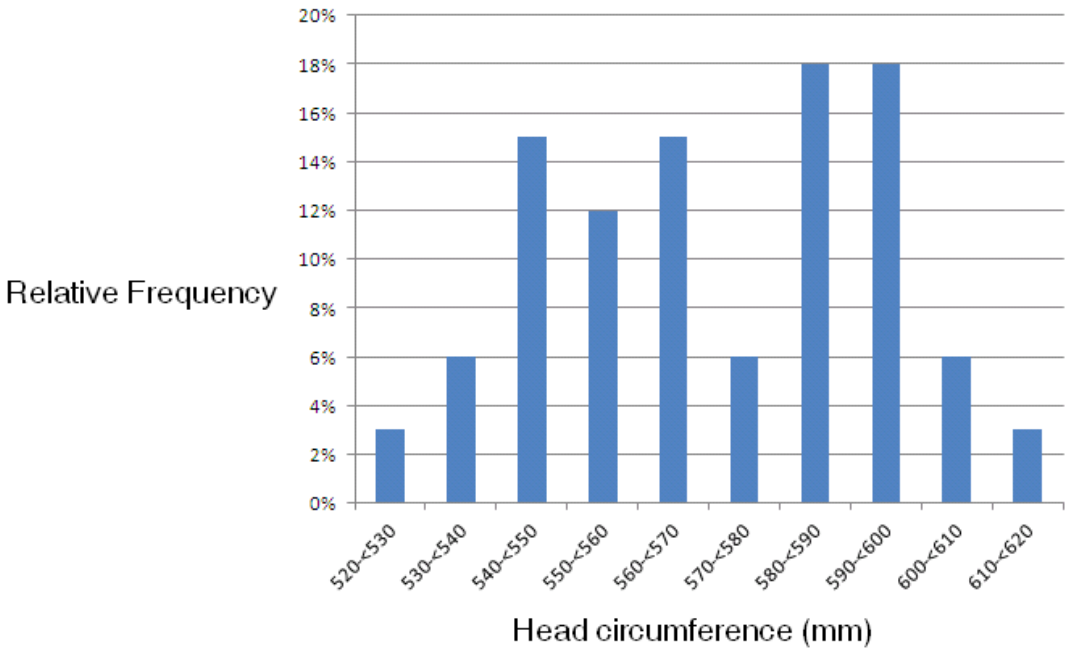
Consider the following:

The frequency and relative frequency for each stem was calculated.

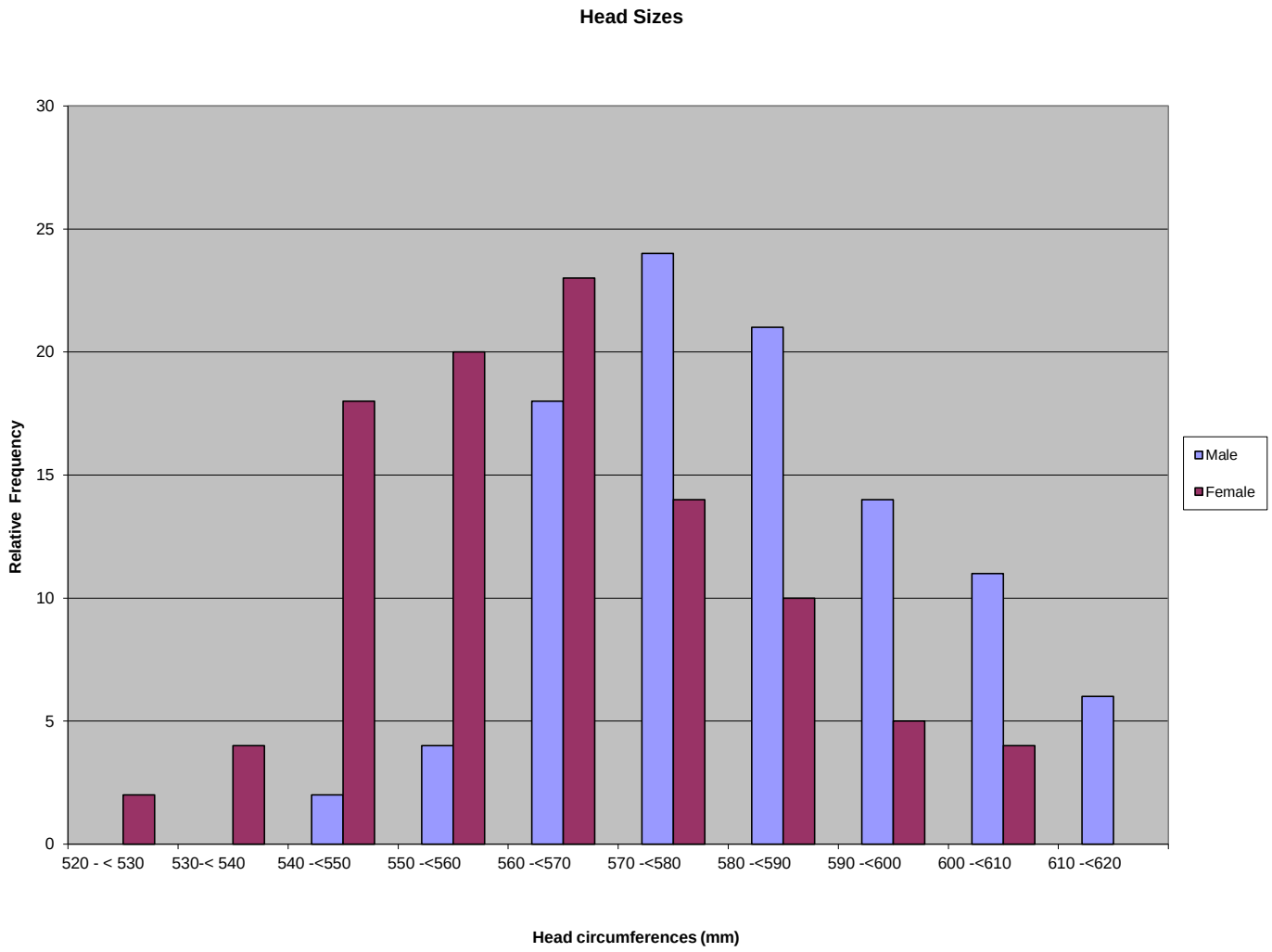
			Frequency	Relative Frequency
52	2	520 - < 530	1	1/34 = 3%
53	1 6	530 - < 540	2	2/34 = 6%
54	0 5 5 5 6	540 - <550	5	5/34 = 15%
55	0 1 2 5	550 - <560	4	4/34 = 12%
56	0 3 4 5 8	560 - <570	5	5/34 = 15%
57	3 7	570 - <580	2	2/34 = 6%
58	0 0 4 5 5 6	580 - <590	6	6/34 = 18%
59	0 0 0 3 4 5	590 - <600	6	6/34 = 18%
60	0 0	600 - <610	2	2/34 = 6%
61	3	610 - <620	1	1/34 = 3%

Using Excel we can draw a histogram. The diagrams below show two representations. Examine the axes. When would it be more suitable to use relative frequency as opposed to frequency ?





Look at the following histogram showing the distribution of head sizes for a different group of males and females. Compare the distributions. Is there any evidence to suggest that there are differences in the head sizes of men and women?



Why do you think the relative frequency is used for this histogram? Does it matter that the actual numbers of males and females in this sample are not given?

Consider the following piece of research.

In the late 1990s a **study** was undertaken in the South Island of New Zealand to explore iron levels in babies and toddlers (age 6-24 months). The participants were selected ***randomly*** from Christchurch, Dunedin and Invercargill (South Island Urban).

The iron, fibre, calcium and vitamin C intake per day was collected over three non-consecutive days and the ***iron (ferritin)*** levels in the blood were measured. Information such as whether the child was being breastfed, fed with formula milk or cows milk, as well as things like gender, ethnicity, maternal education, income level of household, if there were smoker(s) in the household and marital status of the mother were also obtained.

What **type** of study was this? Explain your choice.

Suggest how the participants may have been selected.

Having explored the literature, a number of factors were suggested that could have had an effect on the levels of iron or ferritin in the blood. One of these was gender – boys are at higher risk of having reduced levels of iron or ferritin in their blood.

Some sample data from the study is shown in the tables oopposite.

Ferritin Levels	Sex
22.80	1.00
8.00	1.00
	1.00
	1.00
16.80	1.00
	1.00
22.40	1.00
	1.00
17.30	1.00
	1.00
48.20	1.00
27.80	1.00
12.40	1.00
	1.00
5.30	1.00
	1.00
22.20	1.00
39.10	1.00
16.20	1.00
	1.00
18.70	1.00
38.00	1.00
	1.00
8.60	1.00
13.50	1.00
13.00	1.00
5.20	1.00
19.40	1.00
6.21	1.00
	1.00

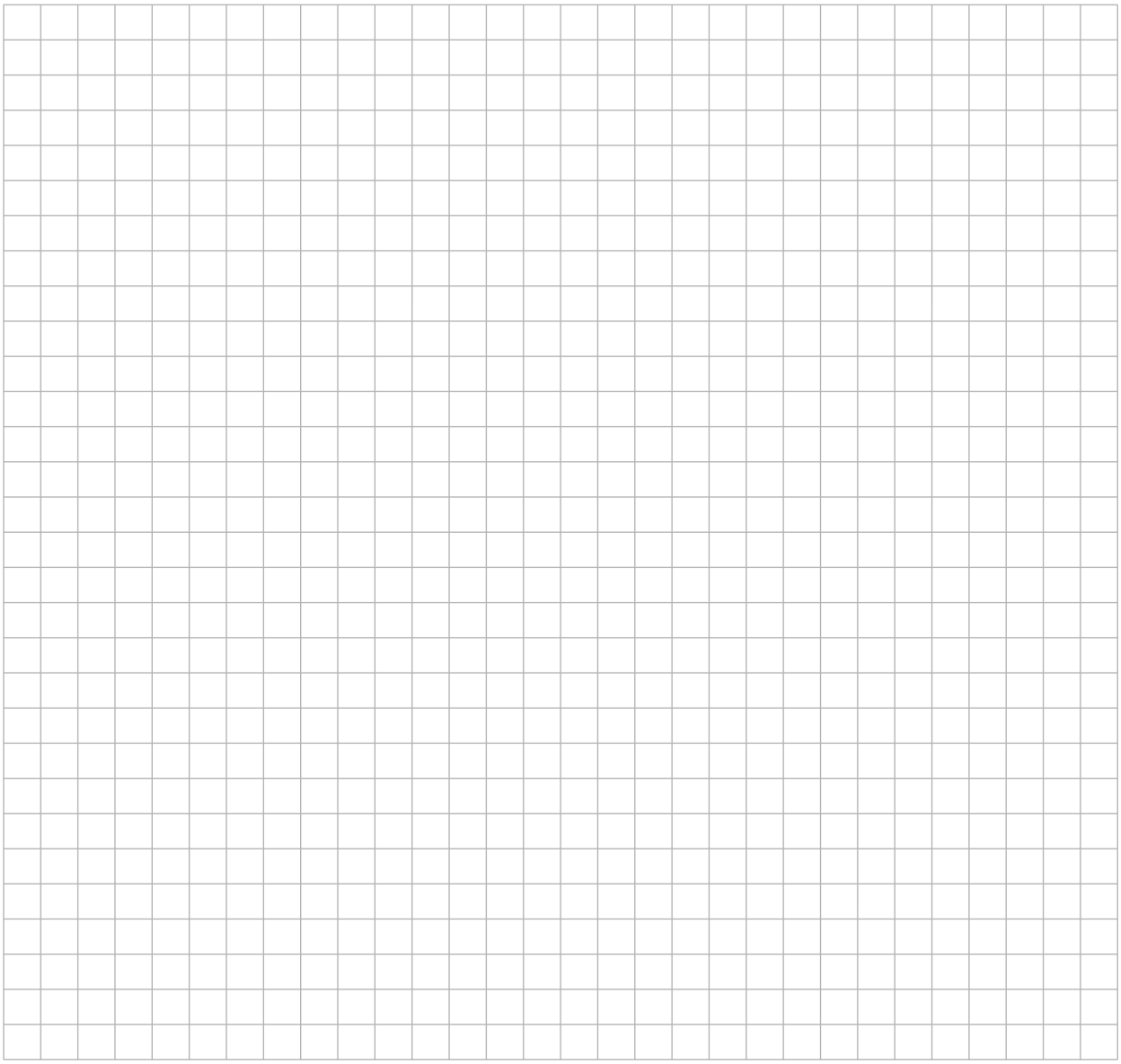
Key

0 = Boy

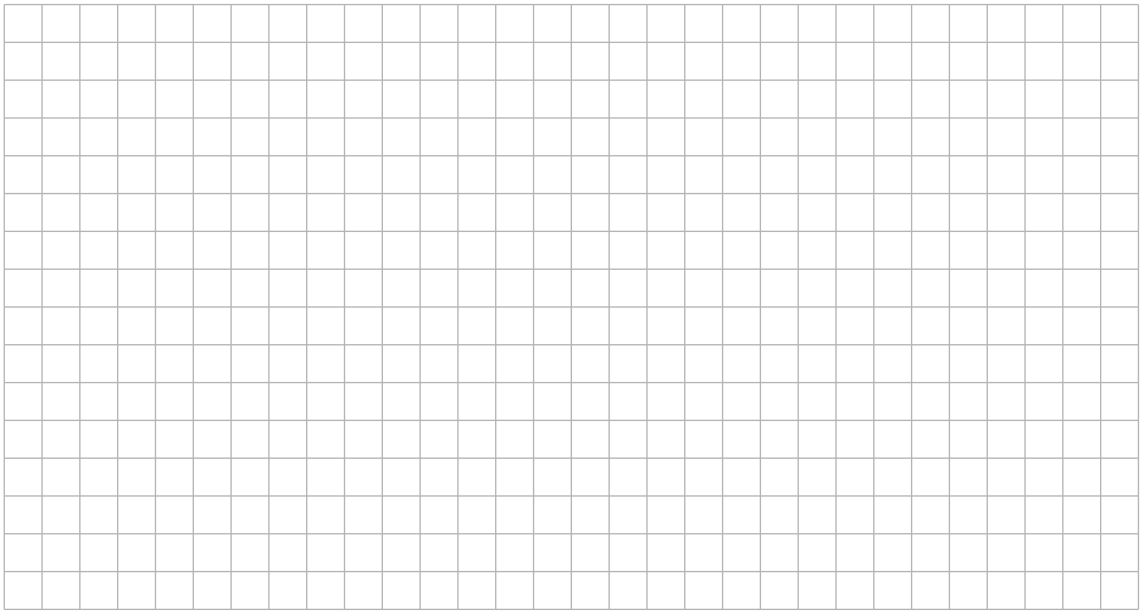
1= Girl

Ferritin levels	Sex
27.20	0.00
	0.00
44.30	0.00
25.00	0.00
17.20	0.00
12.60	0.00
	0.00
9.90	0.00
23.70	0.00
12.75	0.00
	0.00
8.50	0.00
2.00	0.00
32.40	0.00
53.10	0.00
27.50	0.00
16.40	0.00
8.20	0.00
4.80	0.00
16.70	0.00
14.40	0.00
19.50	0.00
20.20	0.00
11.50	0.00
32.70	0.00
	0.00
12.70	0.00
6.30	0.00
10.40	0.00
8.00	0.00

Display this data in a way that will allow you to see patterns in the variation and compare the two **distributions.**



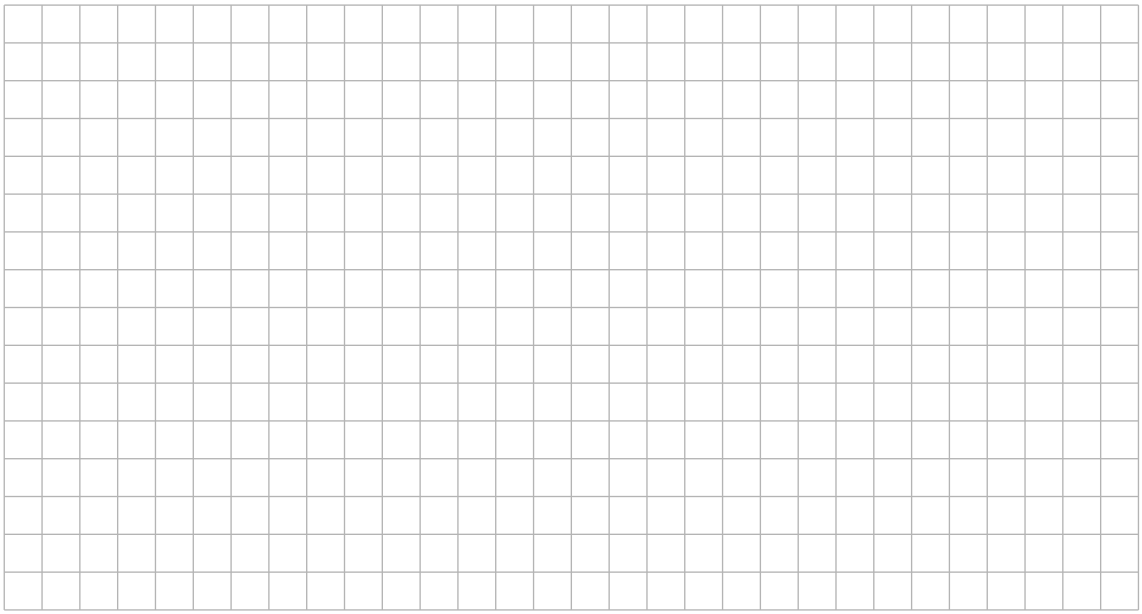
Describe and compare the **shape** of both **sample distributions.**



Describe and compare the **spread** of both **sample distributions.**

Use the data to answer the question “Do the iron levels of South Island urban boys tend to be lower than the iron levels of South Island urban girls?”

Explain why you have made this conclusion.



What aspects of this question did you find confusing?

Was the data difficult to display? What caused this difficulty?

What type of display did you decide to use? How did you make your choice?

How did you deal with the missing data when you displayed your data?

Did the missing data have any impact on the conclusions you made about the study?

Now see how well you have understood these concepts by answering the examination style questions. Make a note of the parts of the questions you find difficult or confusing. You may like to discuss these areas with your friends or your teacher.

Further examination questions can be found on the State Examinations Commission website (www.examinations.ie), including Sample Papers and the June 2010 and 2011 examinations.