Summary (Univariate) investigation for juniors.

(Levels 2 – 6)

Name: _____

Current Level (circle one):

2 / 3 / 4 / 5 / 6

Goal Level (circle one):

2 / 3 / 4 / 5 / 6

Exercises:

1) Do you remember what a summary question is?

Problem

I wonder if what the height of students at Aorere College tends to be?

Plan

We will ask 30 boys and 30 girls the following question:

- How many texts do you send each day?

2) Do you think all students will respond with an accurate answer to this question?

3) Why might students give an incorrect answer to the question?

4) Why is the wording of the following question biased?
"Why is it cool to send texts?"

Plan

Data collection methods

Data can be collection in a number of different ways. Here are some:

- Survey
- Questionnaire
- Poll
- Census

Exercises:

Discuss what each of these methods are, and write down an explanation.

1) A survey is …

2) A questionnaire is …

3) A poll is …

4) A census is …

Data

When you collect your data, you need to record it in a table.

You also have to decide on your **sample size**.

For **count** data:

use a sample size of **50**.

For **measurement** data:

use a sample size of **30**.

Exercises:

1) Circle the words that complete the sentences below.

Smaller sample sizes take a shorter / longer time to collect data, but are more / less accurate.

Larger sample sizes take a shorter / longer time to collect data, and are more / less accurate.

2) What is count data?

3) Why is measuring 50 students when you have count data a good compromise? Explain

4) What is measurement data?

5) Why is measuring 30 students for measurement a good compromise? Explain

6) Why do we have different rules for sample sizes depending on whether it is count or measurement data? Explain

Analysis

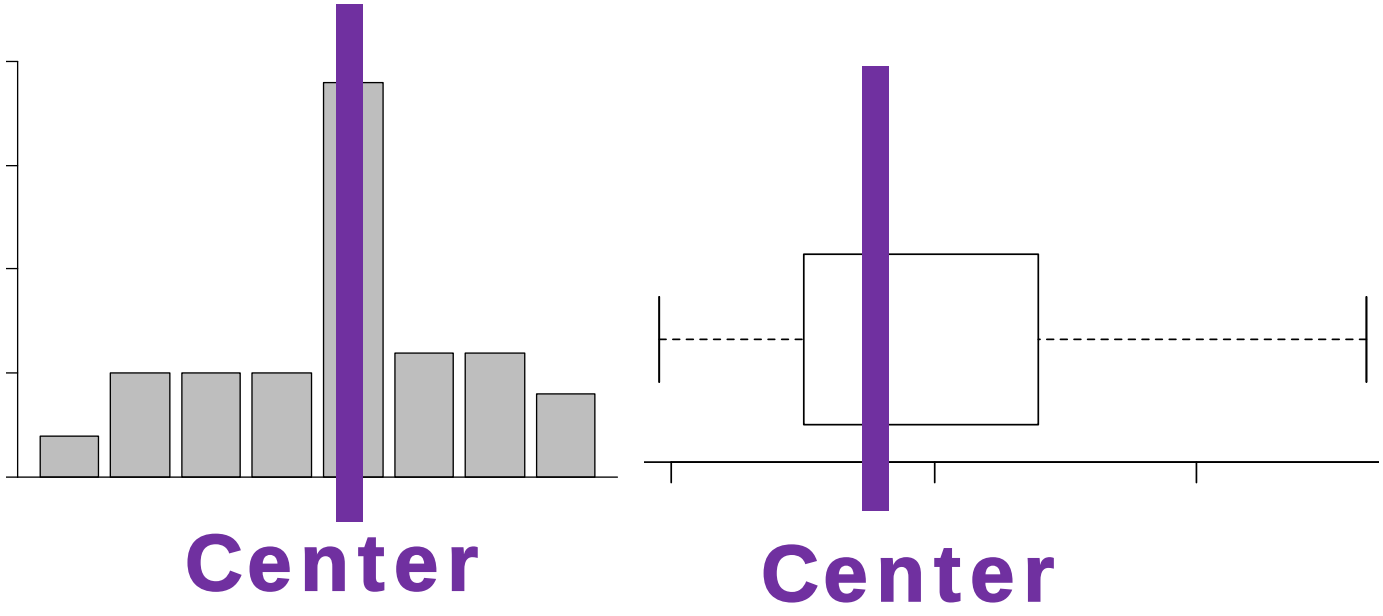
Graphs and Calculations

Sample statistics

Numbers calculated from a *sample* of numerical values that are used to summarise the sample.

The statistics will usually include at least one *measure of center* and at least one *measure of spread*.

Measures of Center



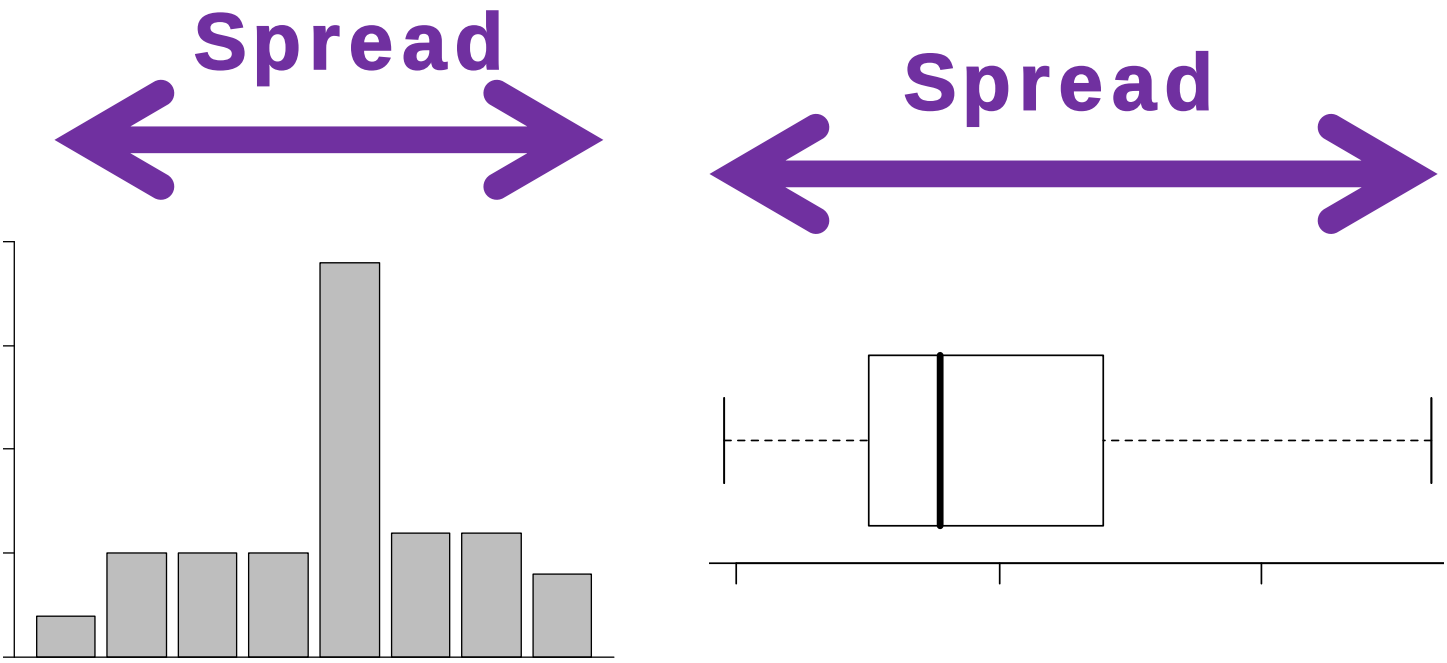
An average is a number which represents the group as a whole. There are 3 measures:

- Mean
- Median
- Mode

Measures of Spread

A measure of spread looks at how precise or accurate the data is. There are two measures you will use:

- Range
- IQR (Inter Quartile Range)



Calculations

With different types of data, different measures of center and spread are used.

We need to know how to calculate each type.

Measures of center

Mean

The mean is calculated by adding the values and then dividing this total by the number of values.

Example

The maximum temperatures, in degrees Celsius (°C), in Auckland for the first 5 days in November 2013 were: 18.6, 19.9, 20.6, 19.9, 17.8,

The **mean** maximum temperature over these 5 days is:

$$\frac{18.6 + 19.9 + 20.6 + 19.9 + 17.8}{5} = 19.36℃$$

Median

The median is the middle value, when the values have been ordered from smallest to largest.

If there is an even number of values, then add the two middle numbers together and divide the total by 2

Example

The maximum temperatures, in degrees Celsius (°C), in Auckland for the first 5 days in November 2013 were: 18.6, 19.9, 20.6, 19.9, 17.8,

Put the values in order from smallest to biggest:

17.8, 18.6, 19.9, 19.9, 20.6

Then find the middle:

17.8, 18.6, 19.9, 19.9, 20.6

The **median** is 19.9°C.

Mode

The mode is the most common value.

Example

The maximum temperatures, in degrees Celsius (°C), in Auckland for the first 5 days in November 2013 were: 18.6, 19.9, 20.6, 19.9, 17.8,

Put the values in order from smallest to biggest:

17.8, 18.6, 19.9, 19.9, 20.6

The **mode** is 19.9°C (because the number occurs twice).

NOTE:

- There can be no mode e.g. in the data set 5, 9, 8, 6, 7 there are no common values, therefore there is no mode
- There can be more than one mode e.g. in the data set 7, 6, 8, 7, 6, 5, 4, 6, 5, 7 there are two common values 7 and 6, therefore they are both modes
- A mode can be 0 e.g. in the data set 0, 1, 2, 0, 0, 3, 2, 0, 1, 0, 3, the most common value is 0, therefore the mode is 0

Exercises:

Calculate the mean, median & mode for the data below:

1) 4, 6, 3, 8, 2, 4, 9

Mean =

Median =

Mode =

2) 4.4 4.7 3.5 2.2 4.2 6.7 2.9 4.4 1.5 2.0 3.3

Mean =

Median =

Mode =

3) 25, 35, 37, 36, 28, 29, 36, 26, 22

Mean =

Median =

Mode =

4) $150, $145, $135, $150, $148, $156, $143

Mean =

Median =

Mode =

Measures of Spread

Range

The maximum is the biggest number.

The minimum is the smallest number.

Range = Maximum – Minimum

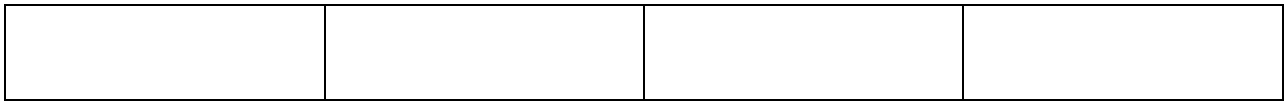
Quartiles

The middle value of a set of data is the median.

Split the data in half, so that half the data is above the median, and half the data is below the median.

The **Lower Quartile (LQ)** is the middle of the bottom half.

The **Upper Quartile (UQ)** is the middle of the top half.



Minimum LQ Median UQ Maximum

Inter Quartile Range

IQR = UQ – LQ

Example

Calculate the range, upper quartile, lower quartile and the inter quartile range of the data set:

6, 6, 6, 7, 8, 8

1) Find the maximum

Max = Biggest number = **8**

2) Find the minimum

Min = Smallest number = **6**

3) Calculate the range

Range = Max – Min = 8 – 6 = 2

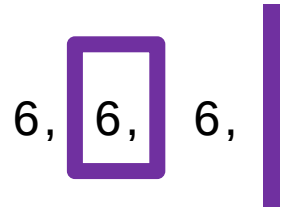
4) Find the median by splitting the data into two halves

6, 6, 6, | 7, 8, 8

5) Take the numbers **below** the median

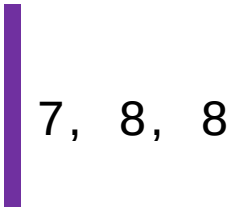
6, 6, 6, |

6) Find the middle number

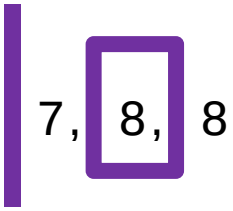


LQ = 6

7) Take the numbers **above** the median



8) Find the middle number



UQ = 8

9) Calculate the inter quartile range

$$\mathbf{IQR = UQ - LQ = 8 - 6 = 2}$$

Exercises:

For each of the following data sets:

- Put the data in order from smallest to biggest
- Find the maximum and minimum
- Calculate the Range,
- Find the Upper Quartile and Lower Quartile
- Calculate the inter-quartile range

1) 35, 48, 36, 24, 19, 56, 43, 23

Range	
Upper Quartile	
Lower Quartile	
IQR	

2) 3.5, 4.2, 2.6, 3.9, 2.8, 3.9, 4.2

Range	
Upper Quartile	
Lower Quartile	
IQR	

3) $45, $35, $56, $29, $89, $76, $83, $74, $21, $42

Range	
Upper Quartile	
Lower Quartile	
IQR	

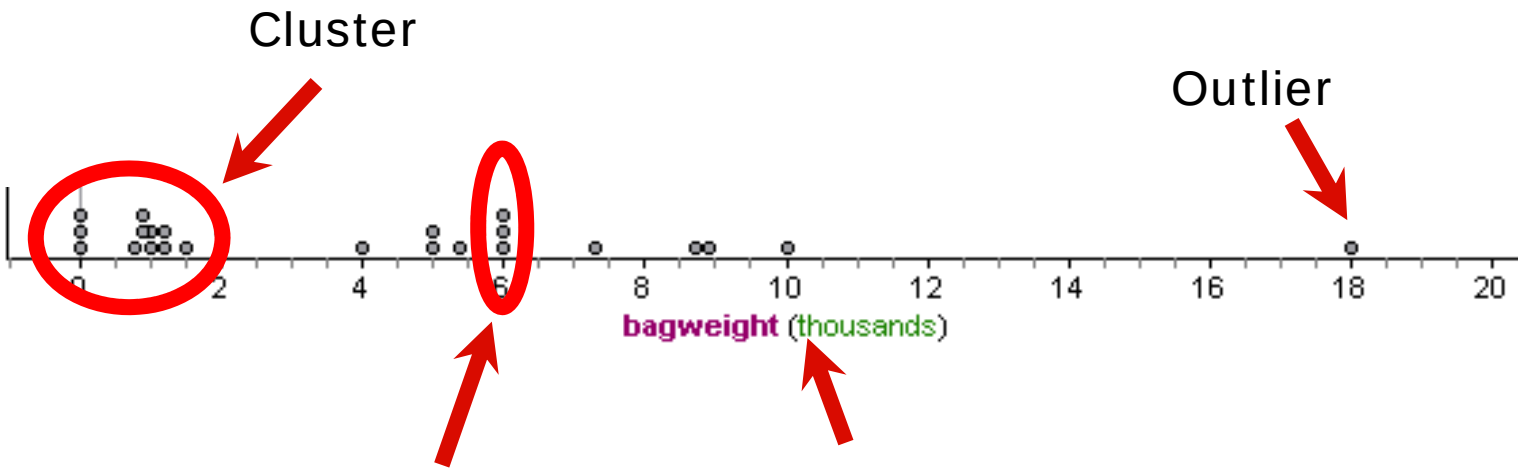
Displaying data

In statistics, it is useful to visualise our data.

This will help us describe the important features of the data such shape, the spread, the center and anything else interesting or unusual (e.g. clusters of data, extreme values or outliers).

The dot plot.

Below is a dot plot of a sample of 30 students from the Census at schools survey for 2011.



The data is stacked on top of one another

Axis labels and units are important

How to draw a dot plot.

Consider this data set of student bag weights from the 2009 Census at schools survey.

	A
1	bagweight
2	900
3	1500
4	1600
5	1900
6	2000
7	2000
8	3000
9	3000
10	3200
11	3200
12	3500
13	3500
14	4300
15	5000
16	5000
17	5000
18	5900
19	7000
20	7000
21	7600
22	9000
23	9600
24	10000
25	16800

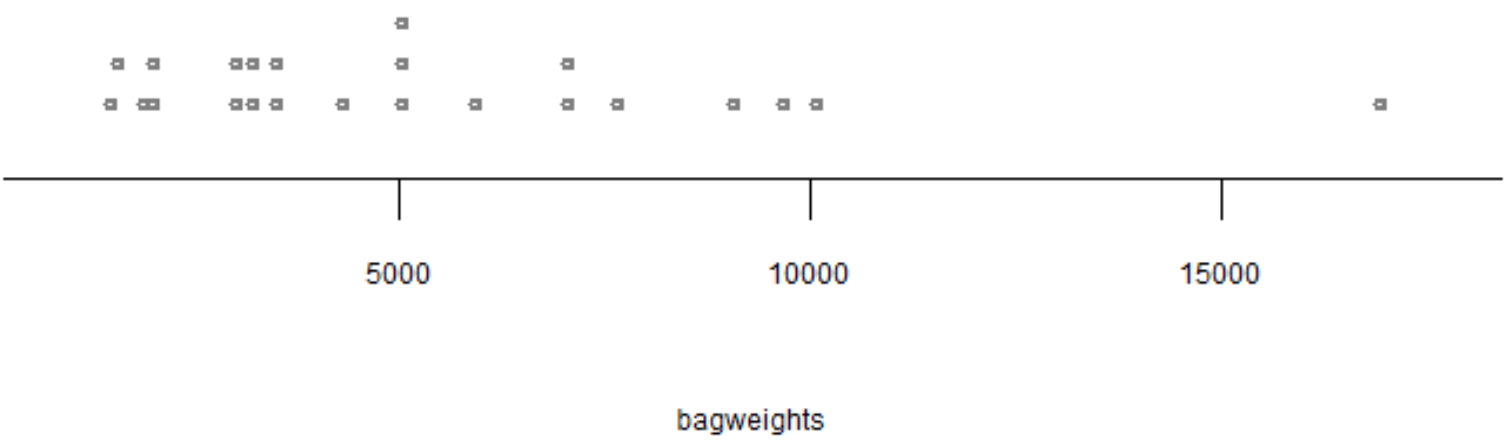
1) Draw an axis for your results. Remember to label your axis

2) Use dots to plot the data points on the axis.

   Note: If a certain weight is repeated, stack the dot on top of the one already there.

3) Finish plotting the remaining data points.

   Note: if you count the number of dots, this should equal the number of data values.



Exercise:

The dataset below shows weekly income of 30 people from the NZ Income Survey.

Weekly Income ($)

120	190	200	240	290	380	456	460	480
480	504	504	504	552	580	590	624	624
630	690	768	800	852	936	1100	1152	1248
1404	1428	1464						

Draw an axis with the appropriate scale and label, and plot the points to create a dot plot of weekly income.

The Stem and Leaf plot.

The **stem** refers to the central (vertical) spine of the data.

The **leaf** refers to the numbers shooting off horizontally from the spine.

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

The above stem and leaf plot shows the weight (kg) of students in a particular university course. Part of the dataset is shown below.

C
Weight
51
55
57
57
60
60
60
60
60
60

How to read a stem and leaf plot

1. Start at the stem. This first number is ‘5’.
2. On the leaf of ‘5’ the first number is ‘1’.
3. The first result is the 51kg.
4. The second result is 55kg.
5. The largest result is 120kg.

Exercise:

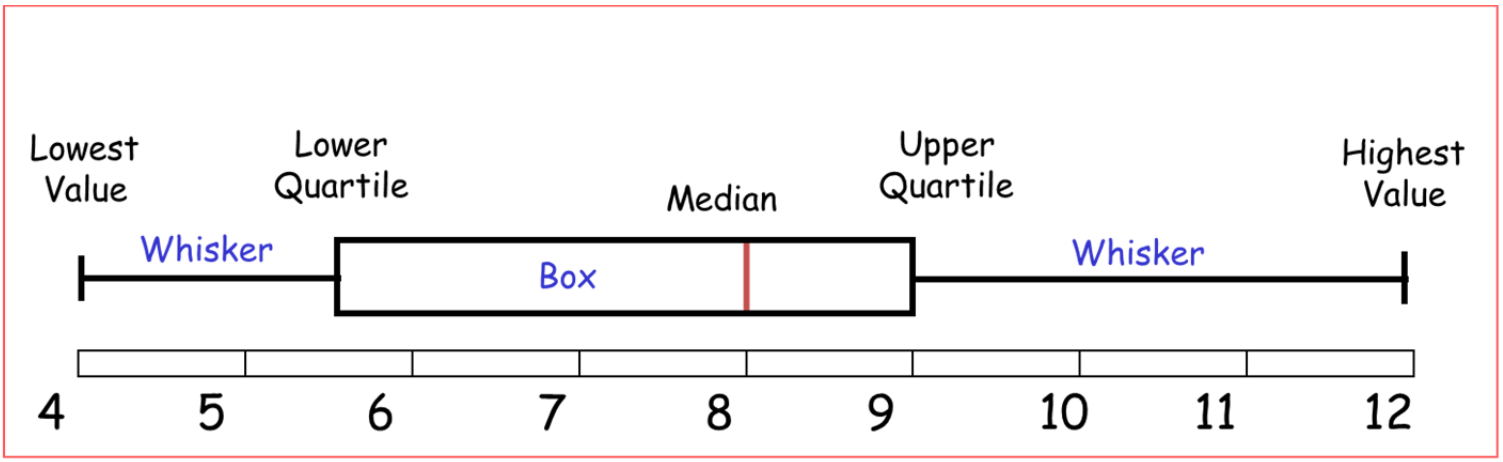
Create a stem and leaf plot of the dataset and answer the questions.

Rubber band lengths (mm)										
296	223	188	256	216	268	214	260	229	205	152
283	225	189	245	211	235	190	165	239	193	237
283	225	177	241	210	238	199	167	233	242	173
270	221	174	247	208	233	191	164	238	246	177
272	270	220	223	200	200	208	206			

The box and whisker plot.

The Box and whisker plot (or just box plot) shows minimum, lower quartile (LQ), median, upper quartile (UQ) and maximum values of a dataset.

The box plot is a useful in showing the *center* of the data (the median) and the *spread* of the data around the median.

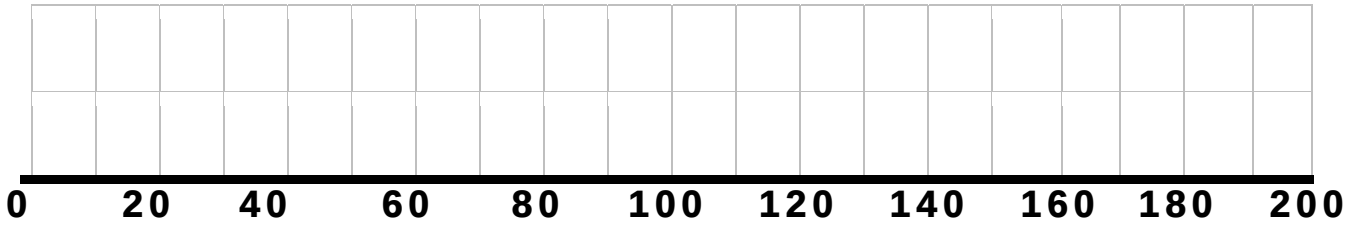


How to draw a box plot.

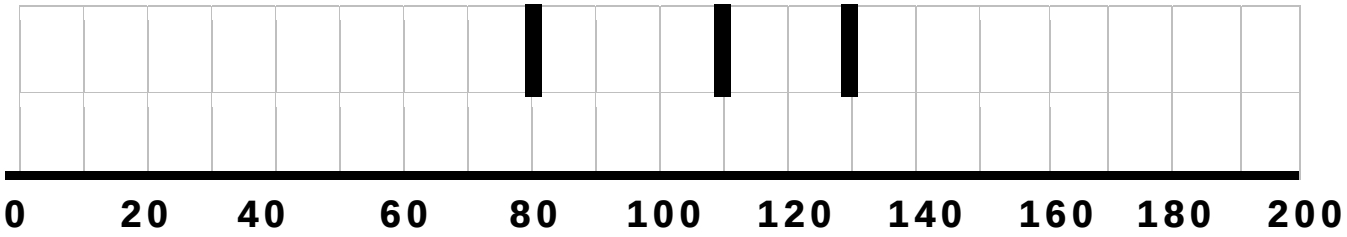
Summary statistics from a make believe dataset are shown below. We will create a box plot from these summary statistics.

Median	=	110
Minimum	=	40
Maximum	=	190
LQ	=	80
UQ	=	130

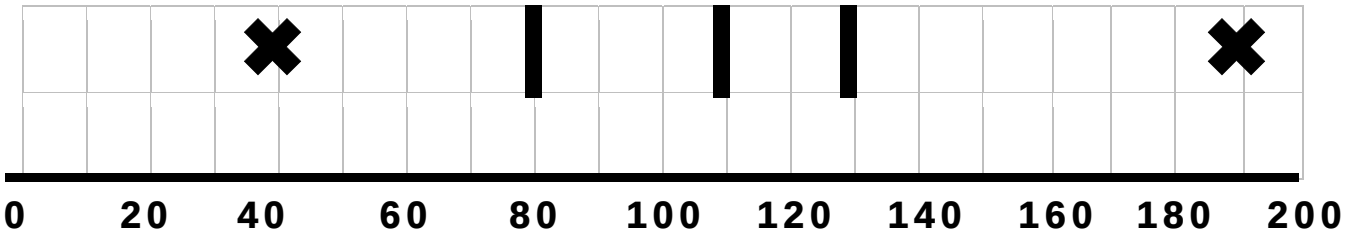
1) Draw an axis for your results



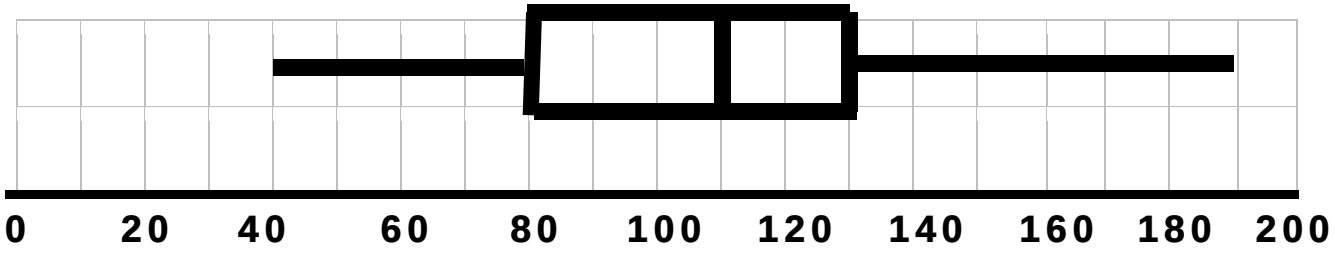
2) Draw the ‘Box’ by plotting the median, LQ, and UQ as vertical lines.



3) Plot the maximum and minimum.



4) Finish off the box plot by drawing lines to close the top and bottom of the box, a line from the minimum to LQ, a line from the max to the UQ.



Exercise:

The summary statistics below are from a sample of 30 people's weekly income from the NZ Income Survey.

Draw an axis with the appropriate scale and label, and plot the points to create a box plot of weekly income.

Clearly label all the important points.

Use the space provided.

Median = $768

Minimum = $456

Maximum = $1,464

LQ = $504

UQ = $1,248

Analysis Interpretation

Here are the features you need to analyse.

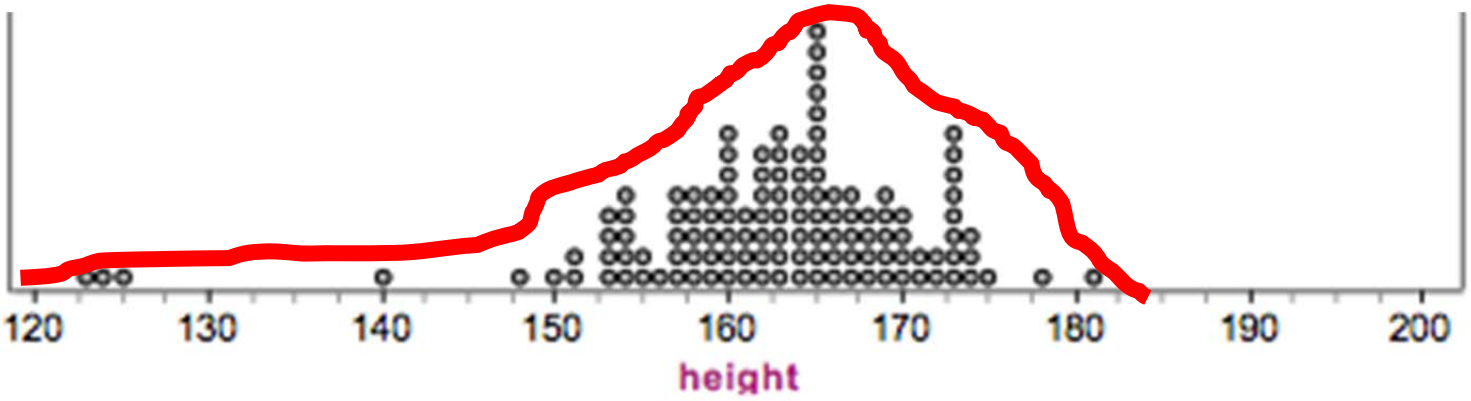
1. Shape
2. Center
3. Spread
4. Middle 50%

We will now go through each feature, before putting it all together.

1. Shape

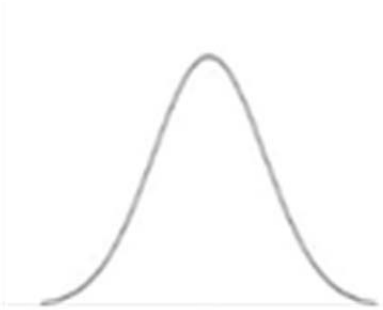
1) The first thing you need to do is sketch a rough shape over the top of a dot plot.

Example

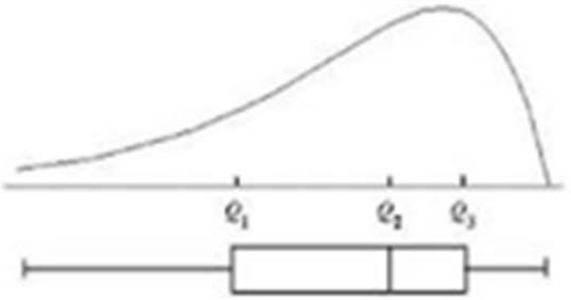


2) The next question you should ask yourself when analysing the shape of the distribution, is “which distribution does my data best match?”

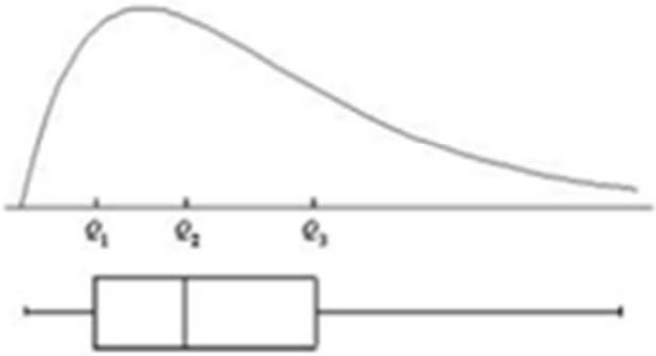
Normal distribution
(Hill/mound shapes, symmetric, Bell shaped curve)



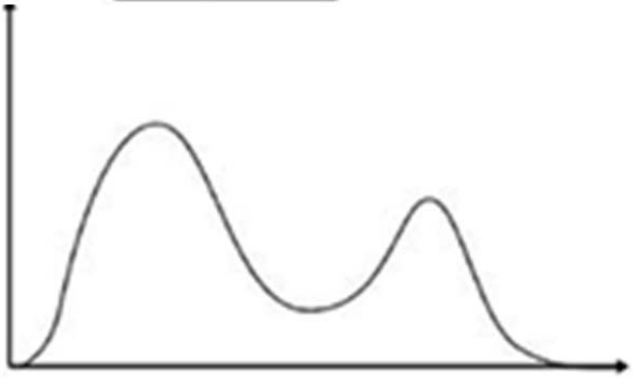
Left skewed
(Tail is on the left hand side)



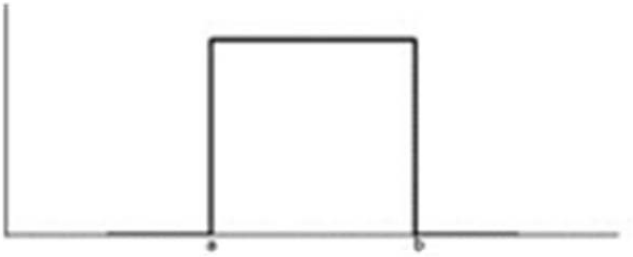
Right Skewed
(Tail is on the right hand side)



Multimodal
(There is more than one peak)



Uniform
(The sides are straight and it looks like a box)



3) Lastly, you need to write a sentence about the shape of each group.

Sentence Framework:

Level 2 / 3:

I notice that the shape of the plot is **hill shaped / a tail on the left / a tail on the right / more than one hill / box shaped** .

Level 4:

I notice that the shape of the distribution is

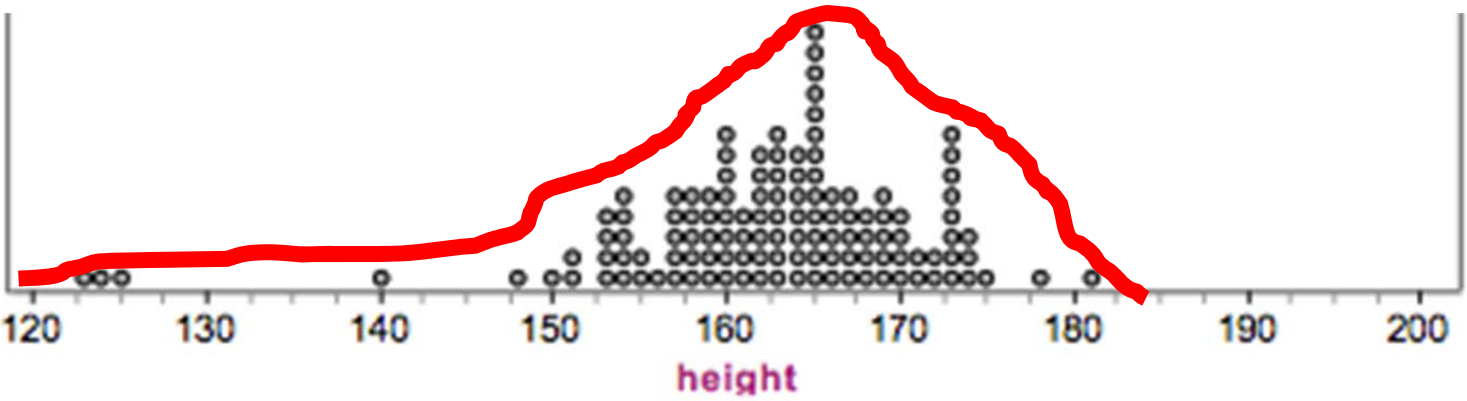
symmetric and bell curve / left skewed / right skewed / bimodal, trimodal, etc / uniform .

Level 5 / 6:

I notice that the shape of the distribution for the

context is **symmetric and bell curve / left skewed / right skewed / bimodal, trimodal, etc / uniform** .

Example



Level 2 / 3:

I notice that the shape of the plot has a long tail on the left.

Level 4:

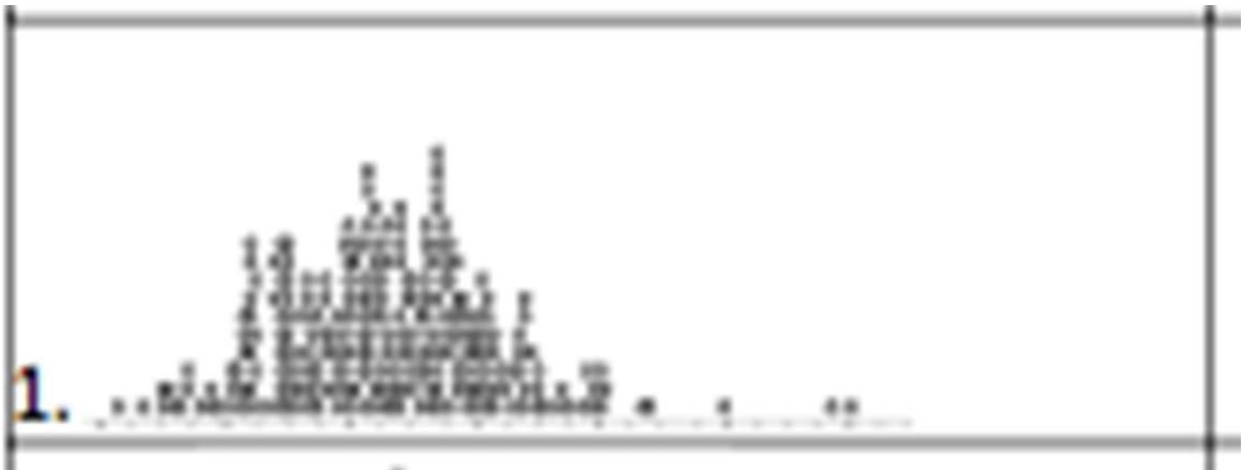
I notice that the shape of the distribution is left skewed.

Level 5 / 6:

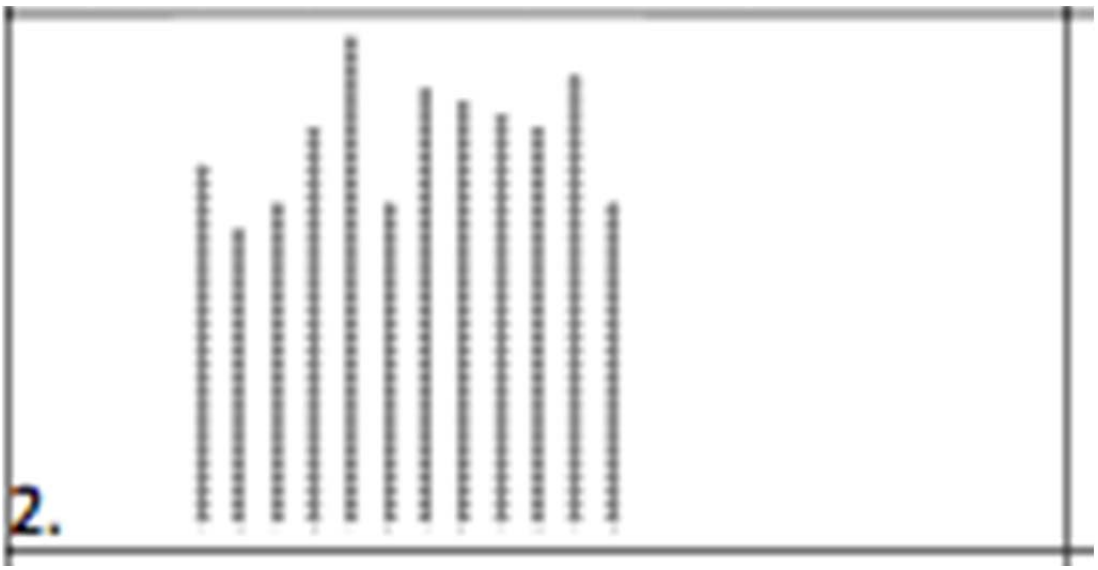
I notice that the shape of the distribution of the heights of students is left skewed as there is a long tail on the left hand side.

Exercise:

Sketch and then describe the shape.

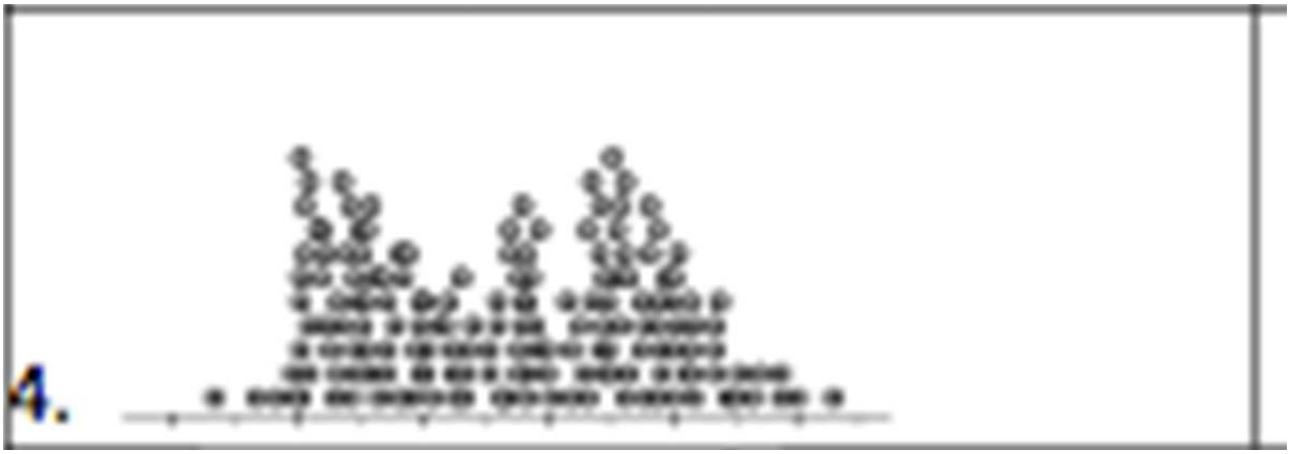


I notice …

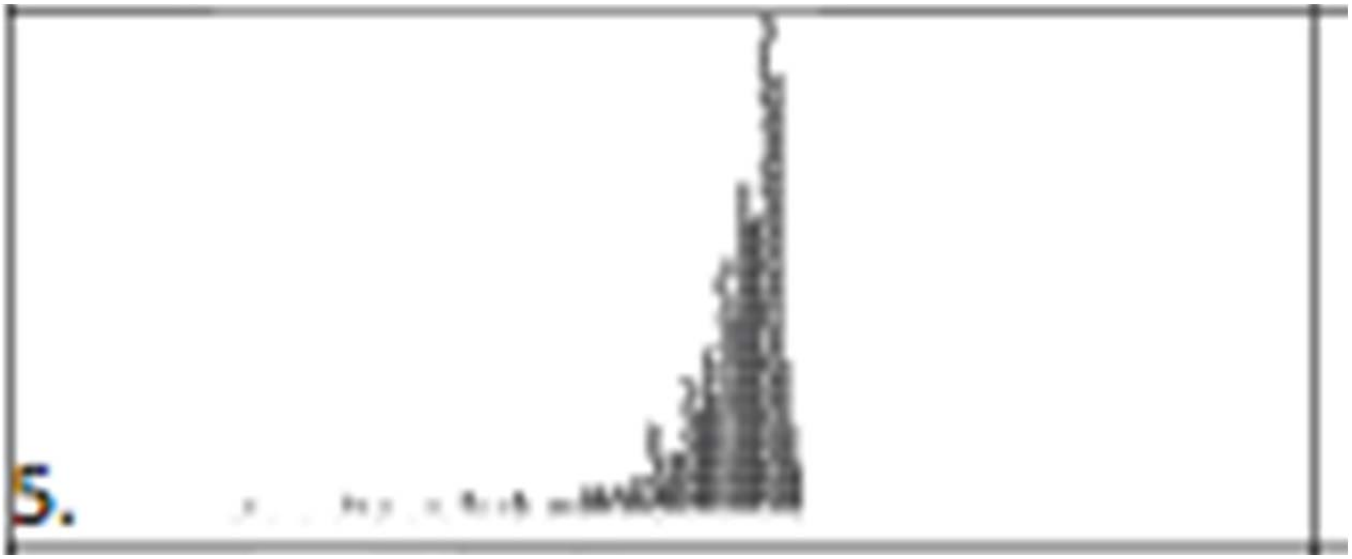


I notice …

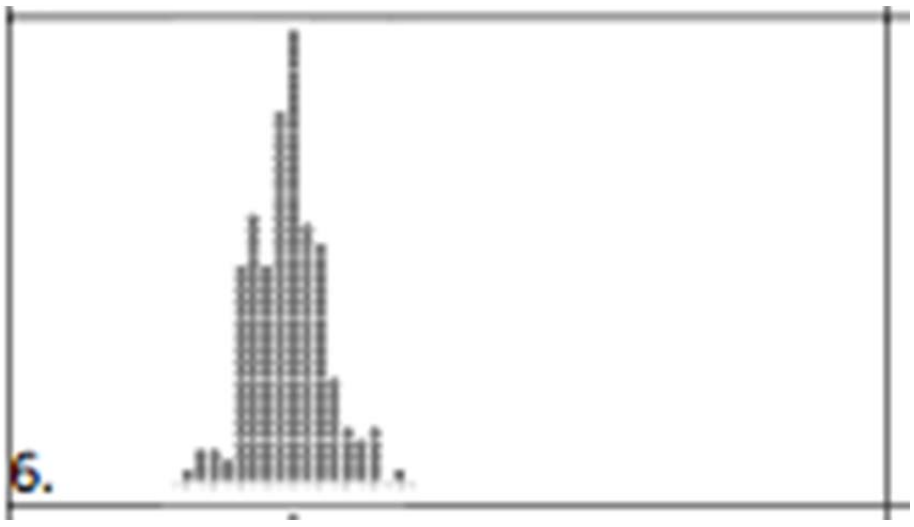
I notice ...



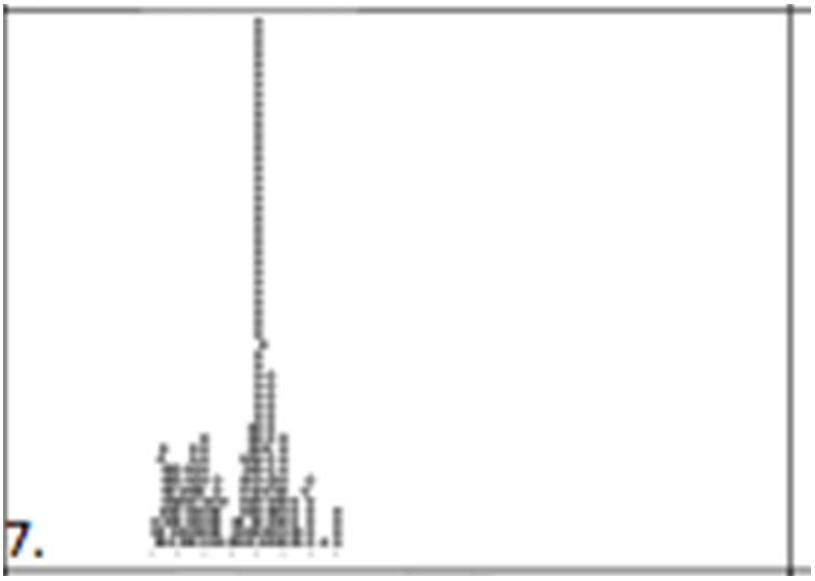
I notice ...



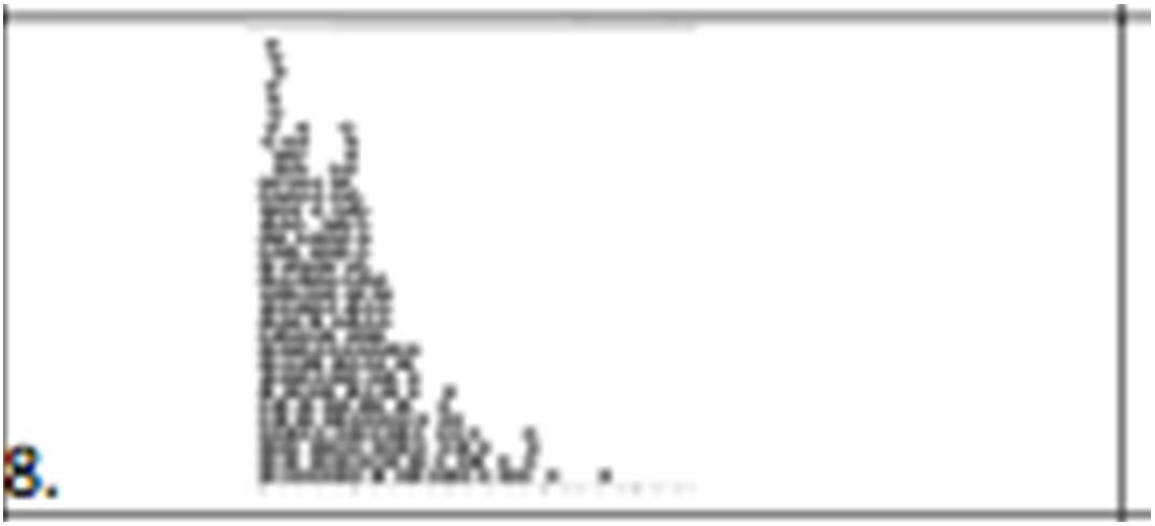
I notice ...



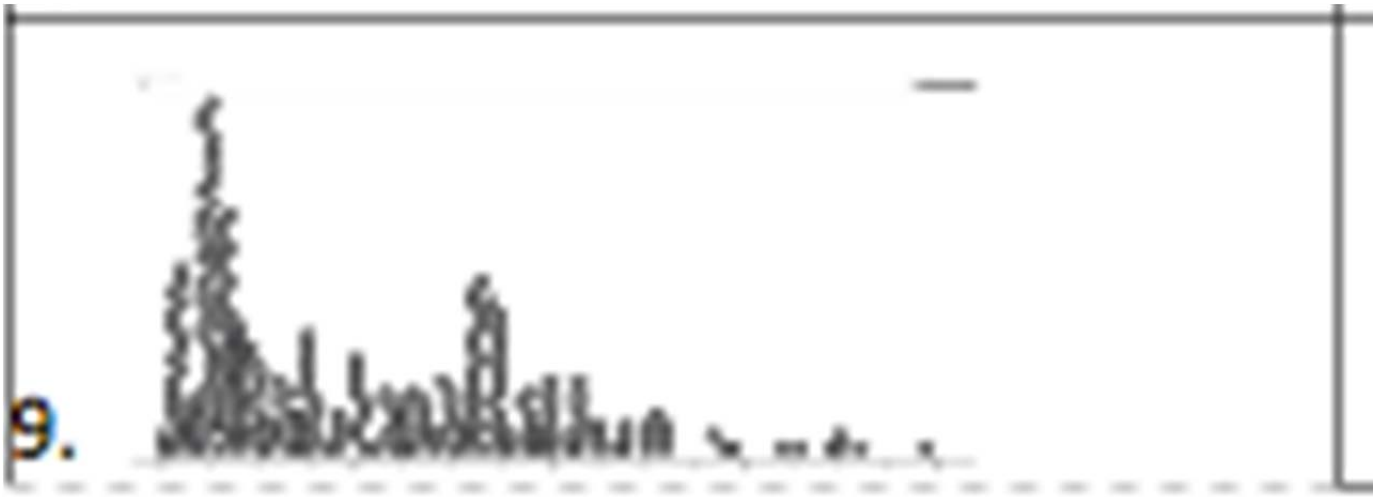
I notice ...



I notice …_____



I notice …_____

9. 

I notice …_____

_____

_____

_____

2. Center

Locate the median, and tell me where it is.

Sentence Framework:

Level 2 / 3:

I notice that the center is ___**median number and units**___.

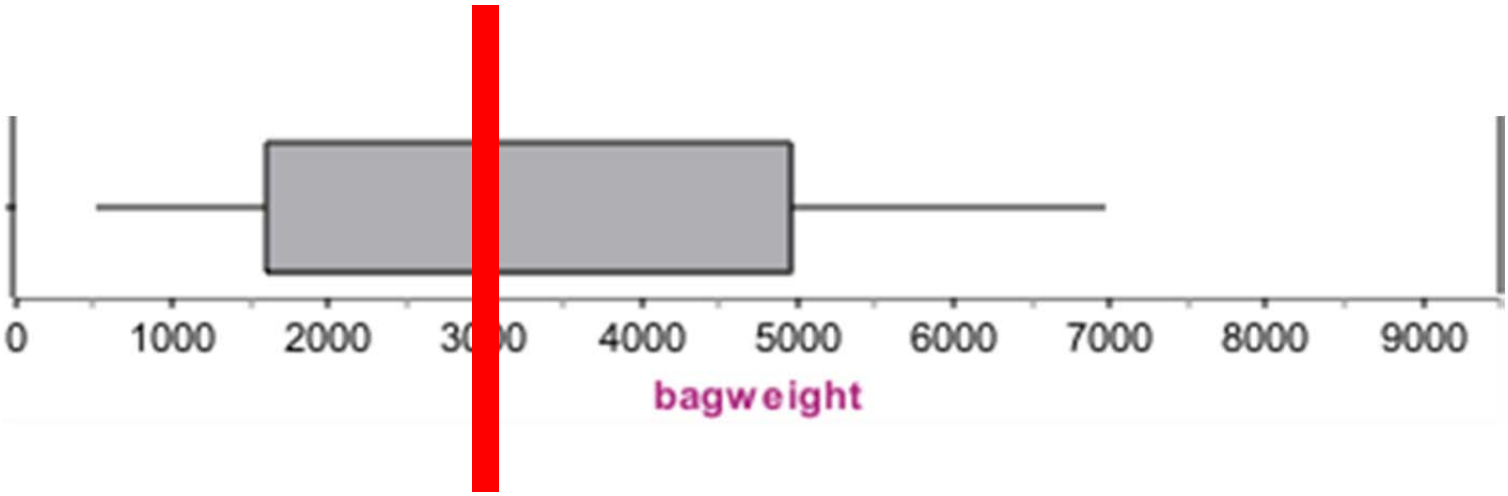
Level 4:

I notice that the median is ___**median number and units**___.

Level 5 / 6:

I notice that the median ___**context**___ is ___**median number and units**___.

Example



Median = 3000 grams

Level 2 / 3:

I notice that the center is 3000 grams.

Level 4:

I notice that the median is 3000 grams.

Level 5 / 6:

I notice that the median bag weight of students is 3000 grams.

3. Spread

Find the Range, and tell me how big it is.

Remember: **Range = Maximum – Minimum**

Sentence Framework:

Level 2 / 3:

I notice that the spread is from **minimum** to **maximum**.

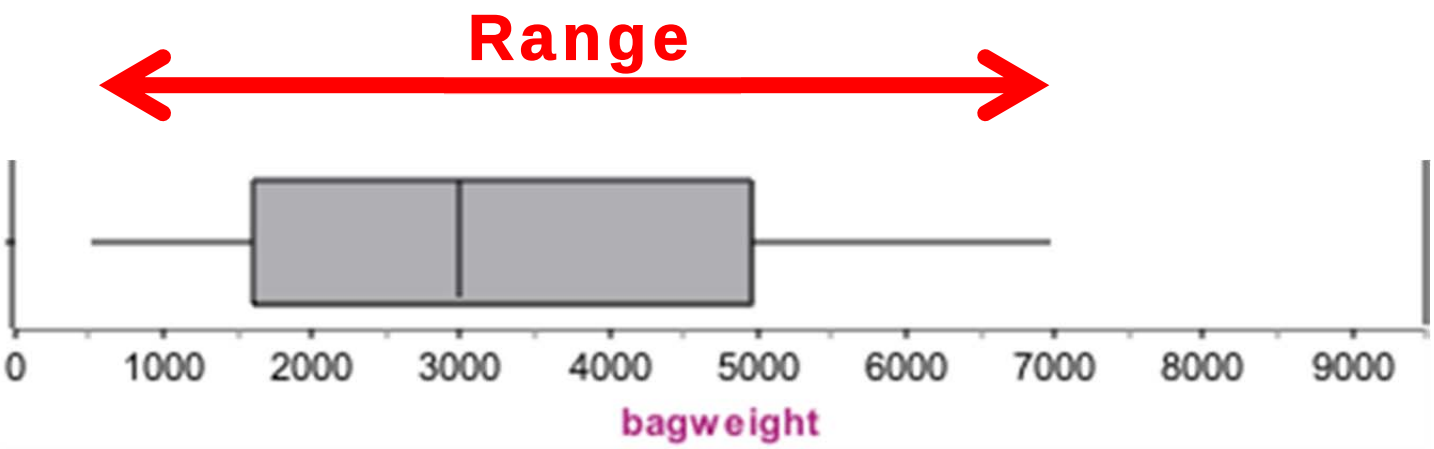
Level 4:

I notice that the range is **range number and units**.

Level 5 / 6:

I notice that the range **context** is **range number and units**.

Example



Range = 7000 – 500 = 6500 grams

Level 2 / 3:

I notice that the spread is from 500 grams to 7000 grams.

Level 4:

I notice that the range is 6500 grams.

Level 5 / 6:

I notice that the range for bag weights of students is 6500 grams.

4. The Middle 50%

We want to see where the middle 50% of the data is located.

The middle 50% of the data are the values that lie between the UQ and LQ.

In other words, the IQR contains the middle 50% of the data.

The easiest way to do this is to look at the boxes on the box and whisker graph (as the box is the middle 50%).

Look at the boxes. Note where the UQ and LQ of each group is.

Sentence Framework:

Level 2 / 3 / 4:

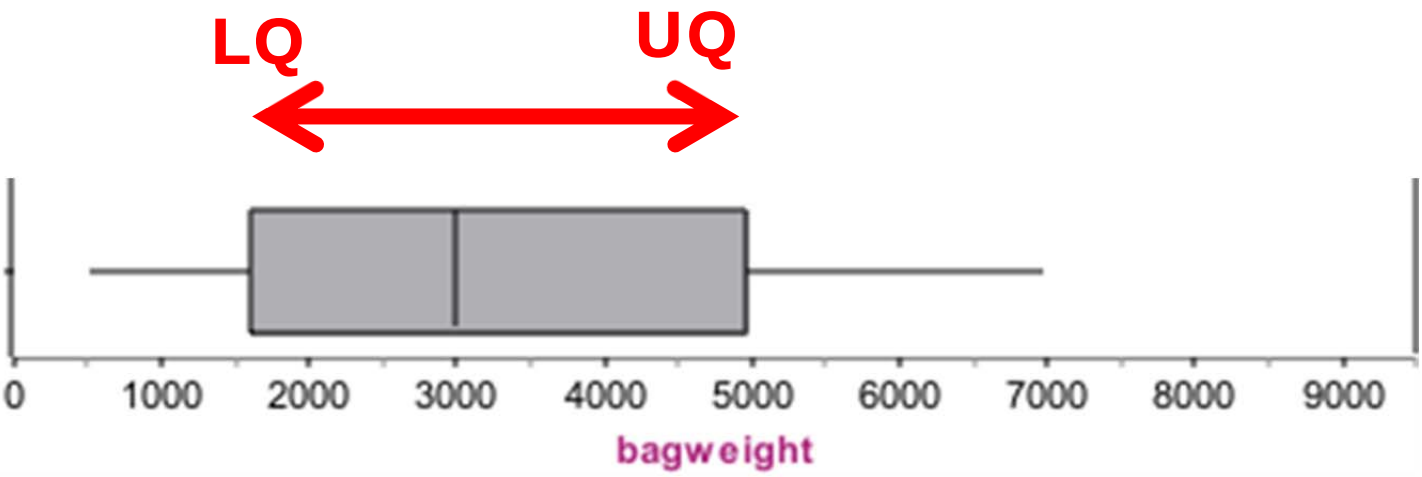
I notice that the middle 50% is from **LQ** to **UQ**.

Level 5 / 6:

I notice that the middle 50% for **context** lies

between **LQ** and **UQ**.

Example



Level 2 / 3 / 4:

I notice that the middle 50% is from 1600 grams to 5000 grams.

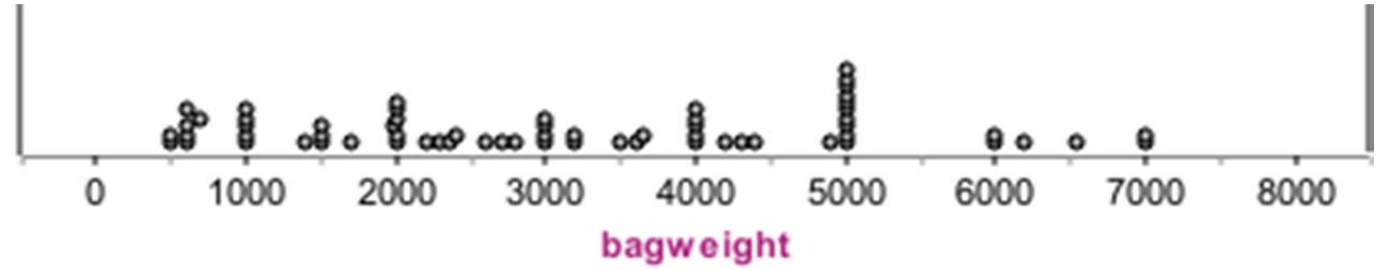
Level 5 / 6:

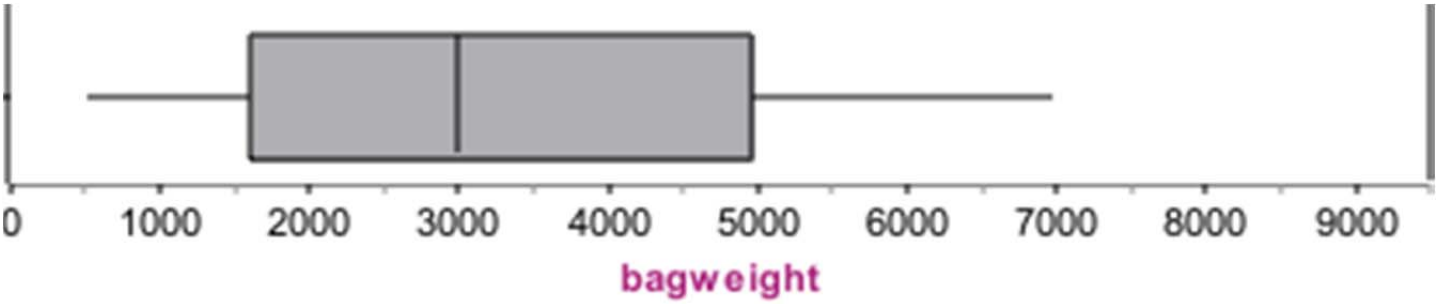
I notice that the middle 50% of student bag weights lies between 1600 grams and 5000 grams.

Example

Analyse the shape, center, spread and middle 50% of the data below.

Problem: I wonder what the weight of school bags of junior students at Aorere College tend to be?





I notice that the shape of the distribution of the bag weights of Aorere College junior students is reasonably symmetric and mound shaped.

I notice that the median of student bag weights of Aorere College Juniors is around 3000 grams.

I notice that the range of the Aorere College Junior student bag weights is around 6500 grams. (Range = 7000 – 500 = 6500grams)

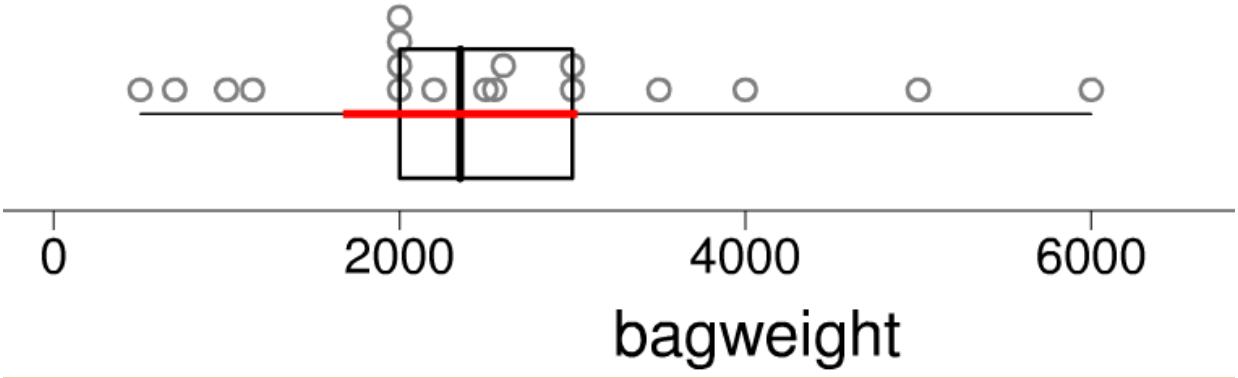
I notice that the middle 50% of Aorere College Junior student bag weights is between 1600 grams and 4900 grams.

Exercise:

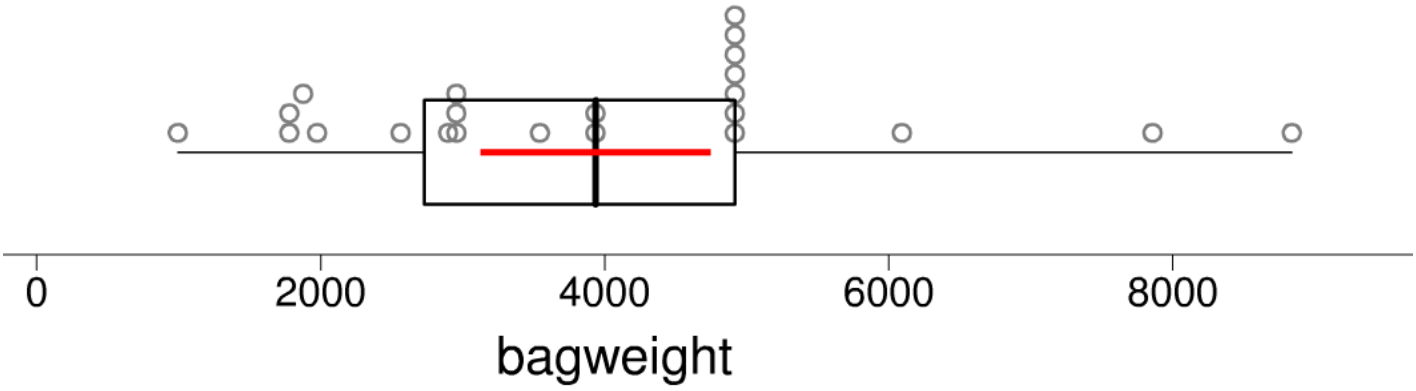
Analyse the following data.

1) How many months old are Aorere College Junior students' cell phones?

2) How heavy are school bags of year 7 Kedgley Intermediate students?



3) How heavy are school bags of year 9 Aorere College students?



Conclusion

There are three things you need to do in your conclusion:

1) Answer the investigation problem
2) Make an inference about the population
3) Discuss Sampling Variability

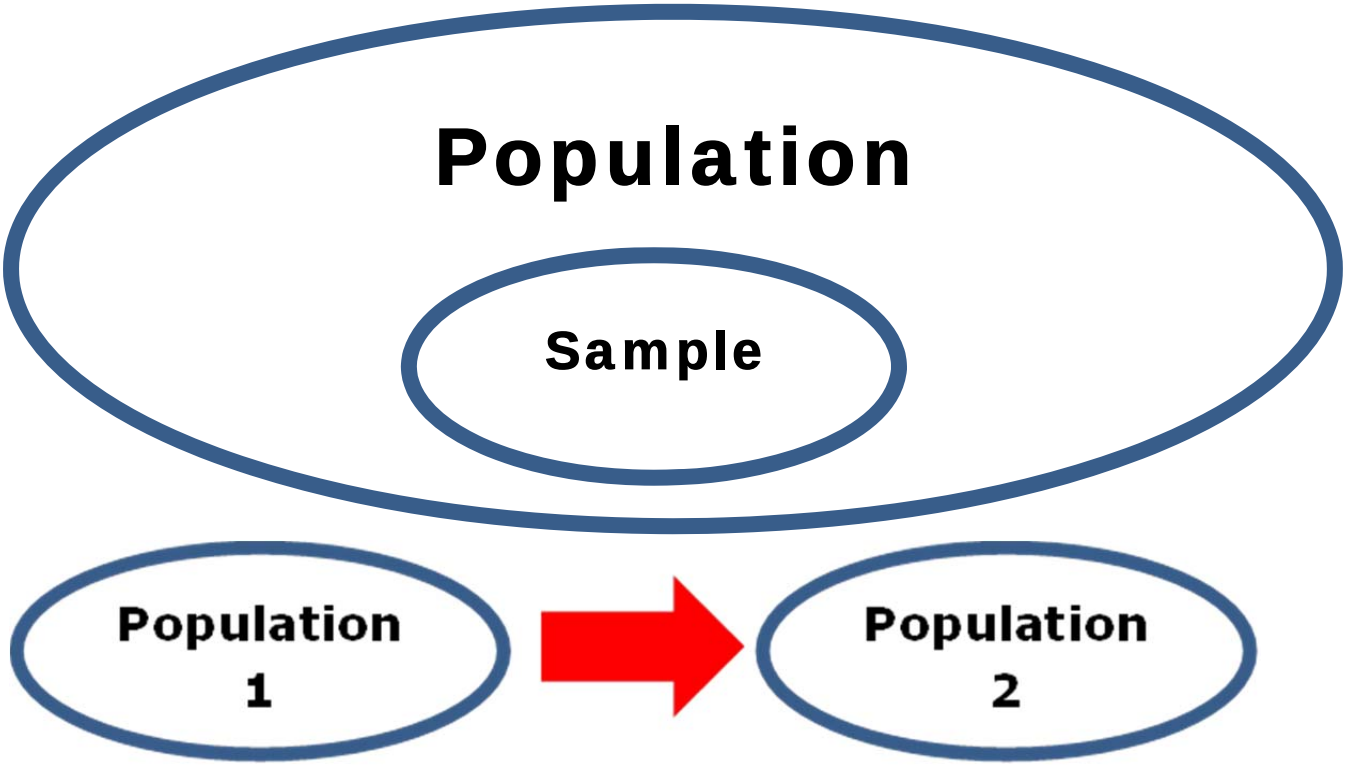
1. Answer the investigation problem

Your problem is usually in the form:

I wonder …

You need to make sure that you answer this question.

2. Inference



The conclusion is valid for the specific population that has been sampled.

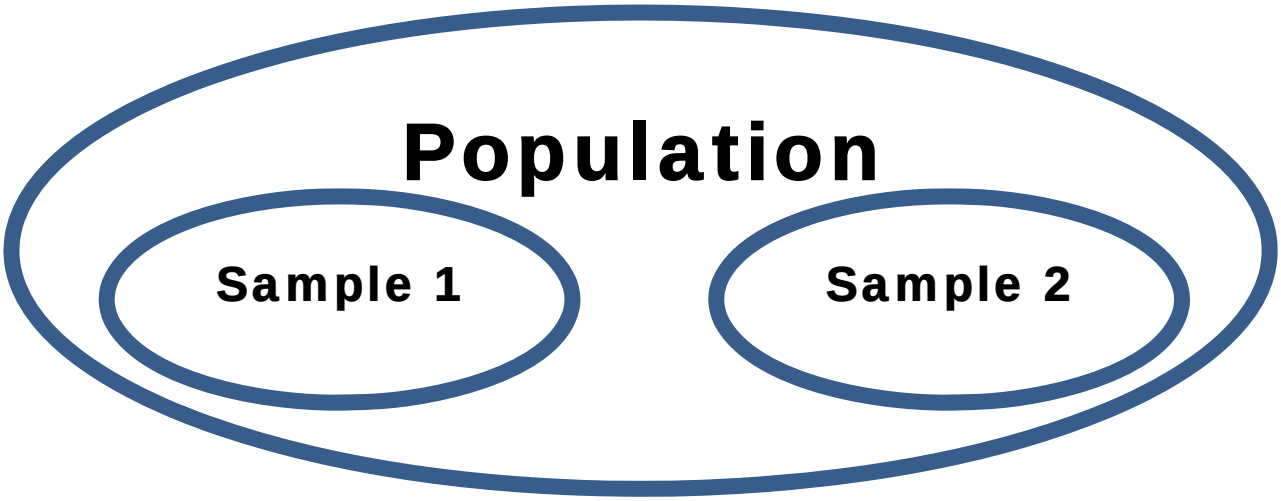
The conclusion may be applied to a similar population to the one for which the data was collected.

Example

If the data is about students in New Zealand, then the conclusions can be applied to New Zealand.

It may be that there are sufficient similarities in the population of Australia and NZ for the data to be useful to help offer guidance.

3. Sampling variability



If I took another sample …

- When another sample is taken, I will select different people, therefore my data will differ from sample to sample.
- Each sample I take should represent the population.
- This means that the analysis and conclusion are likely to remain the same.

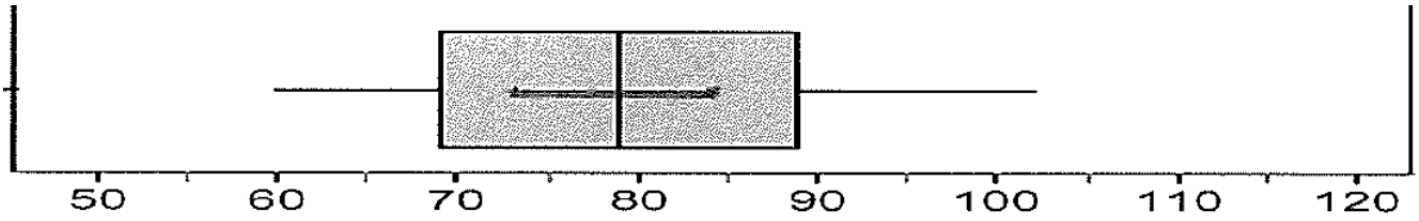
If I took a bigger sample …

- The data will be more representative of the population.
- The results will be more accurate.
- The conclusion will be more accurate.

Example

Problem:

I wonder how long an Aorere College Junior students' ring finger tends to be?



Conclusion:

From our sample, we can conclude that an Aorere College junior student's ring finger tends to be between 70 and 89 cm.

This sample was taken from year 9 and 10 students at Aorere College in Auckland. Therefore it is reasonable that the conclusion can be applied to other Year 9 and 10 students across New Zealand.

However, the conclusion may not be very accurate for students who are older or younger than this.

If I took another sample, I would expect my data to be different, but my results should be similar.

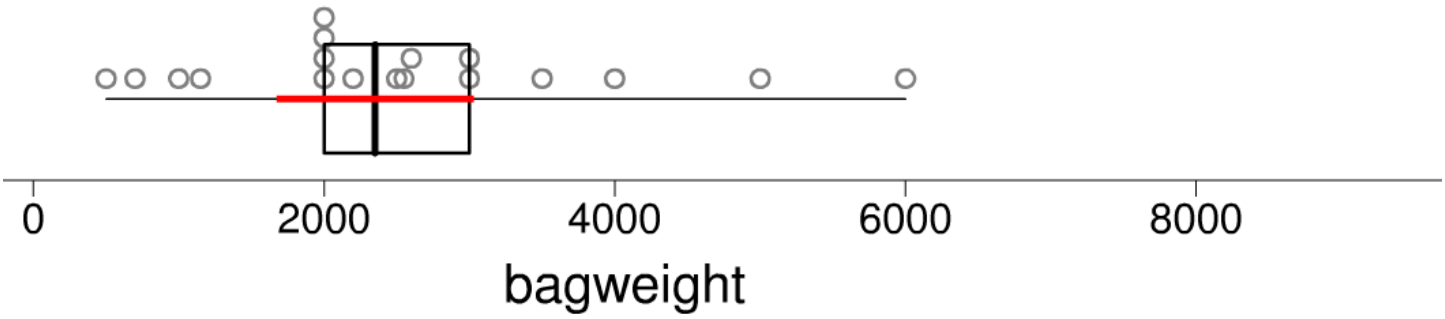
If I took a bigger sample, I would have more information, a more representative sample, and my conclusion would be more accurate.

Exercise:

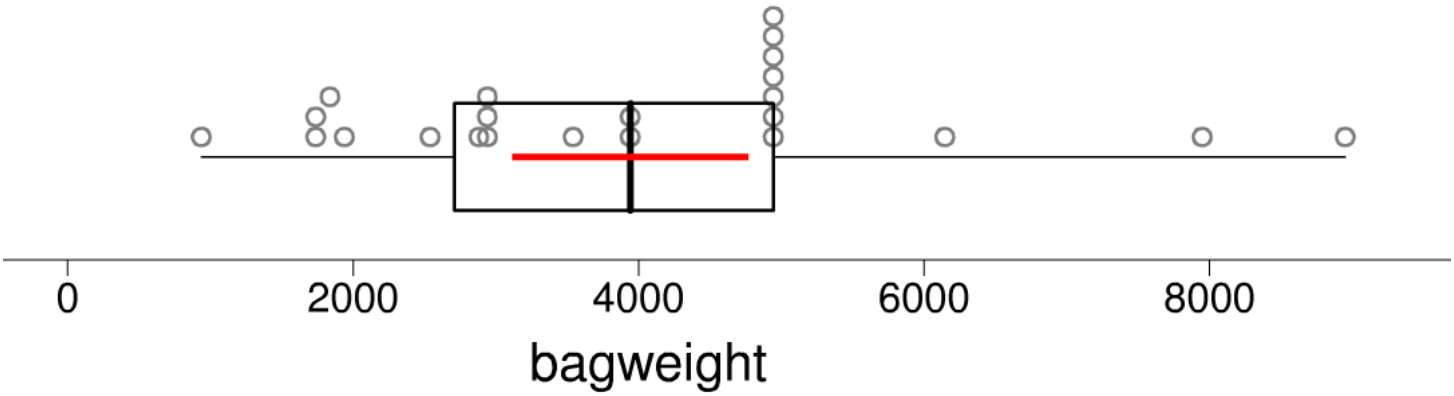
Draw conclusions about the following data.

1) **Problem**: How many months old are Aorere College Junior student's cell phones?

2) **Problem:** How heavy are school bags of Kedgey Intermediate year 7 students?



3) **Problem:** How heavy are school bags of year 9 students?



4) Complete the following sentence:

The population of our school includes _____

5) Would a sample of 30 students from Aorere be representative of the population of NZ? Explain why/why not.

6) Why don't we sample the entire population?

7) It is important for our samples to be randomly selected. Why?

Investigation Exercise

Problem

I wonder how long a Year 9 or 10 student from Aorere College tends to sleep at night?

Plan

Create a detailed plan of how to collect this information.

Data

Create a data table to record the information.

Samples	Hours of sleep per night	Samples	Hours of sleep per night
1		16	
2		17	
3		18	
4		19	
5		20	
6		21	
7		22	
8		23	
9		24	
10		25	
11		26	
12		27	
13		28	
14		29	
15		30	

Analysis

Calculate summary statistics.

Mean	
Median	
Mode	
Range	
Upper Quartile	
Lower Quartile	
IQR	

Draw a dot plot and box and whisker plot.

Analysis

Describe and justify the features of the data. Features may include the center, spread, shape and middle 50%.

Conclusion Answer the investigation question.

Make an inference.

Discussion of sampling variability (taking another sample, changing sample size, etc.)