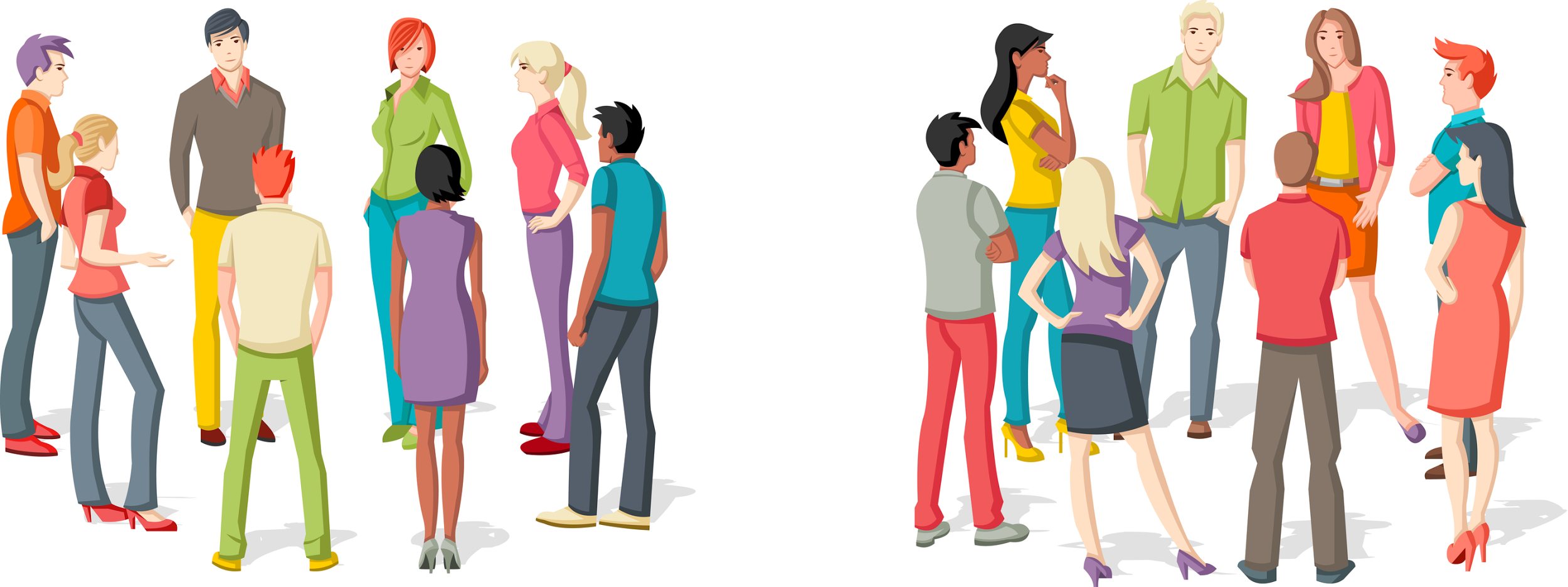


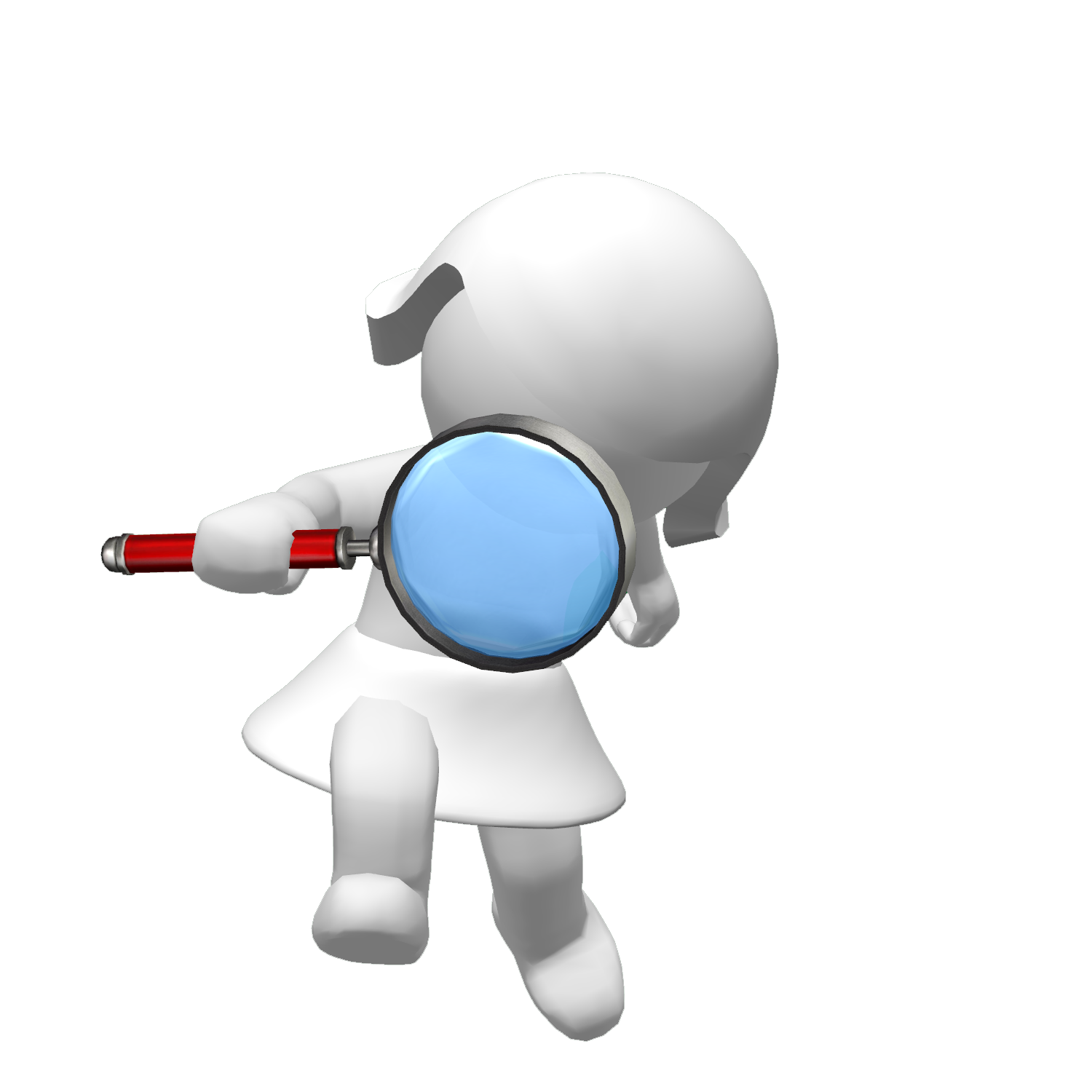
Level 1

Multivariate

Workbook



## Name:



**By Liz Sneddon**

# Problem

Writing Investigation Questions

A comparison question needs:

* Qualitative (2 groups)
* Quantitative (Measurement or Count)
* “tends” (refers to central tendency)
* Direction
* Units

### Example:

I wonder if students who catch the bus **tend** to take longer (minutes) to travel to school than students who walk, for high school students in NZ, for data from Census at School in 2015.

### Exercise:

1. In the example given above:
   1. What are the two groups that are being compared?

### 

* 1. What is the quantitative (measurement or count) variable? And its units?

### 

* 1. What is the population?

### 

### 

* 1. What is the direction?

### 

2) Look at the following questions, and decide if they contain all the information that is needed for the standard. If the question is suitable, then describe the variables (quantitative and group), and the population.

a) Are 8 year old boys generally taller than 8 year old girls in NZ?

### 

### 

### 

b) Do 18 year old males tend to have a longer right foot than 18 year old females in NZ?

### 

### 

### 

c) Are 13 year old boys more untidy than 13 year old girls?

### 

### 

### 

d) I wonder if girls with longer arm spans tend to be taller, in NZ?

### 

### 

### 

e) How does the number of text messages teenage girls send daily compare with the number of text messages teenage boys send daily, in Auckland?

### 

### 

### 

3) Look at the data provided by NZ Census at School (a survey of high school students across NZ) and generate as many different comparison questions as you can on the next page.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Student | Gender | Age | Height  (cm) | Length of arm span  (cm) | *Main* way of travel to school\* | Time taken to get to school (min) | Did *most* at lunchtime\* |
| 1 | male | 12 | 163 | 163 | walk | 10 | Ran |
| 2 | female | 14 | 155 | 155 | bus | 15 | Sat |
| 3 | female | 12 | 155 | 155 | walk | 20 | Ran |
| 4 | male | 10 | 141 | 144 | motor | 6 | Ran |
| 5 | female | 14 | 163 | 164 | motor | 25 | Walked |
| 6 | male | 9 | 144 | 144 | bus | 34 | Walked |
| 7 | female | 13 | 164 | 165 | bus | 37 | Sat |
| 8 | female | 14 | 158 | 118 | motor | 12 | Sat |
| 9 | female | 14 | 166 | 162 | bus | 18 | Sat |
| 10 | female | 10 | 143 | 138 | motor | 14 | Walked |
| 11 | male | 11 | 149 | 144 | bike | 8 | Ran |
| 12 | female | 9 | 140 | 140 | motor | 16 | Ran |
| 13 | male | 9 | 127 | 128 | walk | 19 | Ran |
| 14 | male | 13 | 163 | 163 | motor | 11 | Ran |
| 15 | female | 13 | 150 | 147 | walk | 17 | Ran |
| 16 | male | 11 | 146 | 125 | bike | 7 | Ran |
| 17 | male | 13 | 165 | 154 | motor | 6 | Walked |
| 18 | female | 12 | 159 | 159 | motor | 3 | Walked |
| 19 | female | 15 | 160 | 156 | walk | 56 | Stood |
| 20 | male | 13 | 168 | 175 | walk | 7 | Ran |
| 21 | female | 15 | 170 | 175 | motor | 8 | Sat |
| 22 | female | 9 | 132 | 130 | motor | 5 | Ran |
| 23 | male | 14 | 174 | 182 | motor | 8 | Ran |
| 24 | female | 12 | 150 | 150 | bus | 45 | Stood |

\* Questionnaire wording

Main way to travel to school options: walk, motor vehicle, bus, bike, other.

What you did most at lunchtime options: sat down, stood around, walked around, ran around or played



# Plan

For your assessment, you will need to take a random sample from the dataset. (The reason is so that we can make sure each student has their own dataset to use for the assessment.)

|  |  |  |
| --- | --- | --- |
| **Step 1:**  Click on the button “**Sample and more**”  Screen Shot 2018-01-03 at 11.40.12 AM.png | **Step 2:**  Select the “**Random**” menu. | **Step 3:**  Click on the drop down menu “**Sample with:**”. |
| **Step 4:**  Click on the **group variable** (In this example I selected “Gender”) | **Step 5:**  Enter the **sample size** for **each group** that you want to take a sample from. (Use the rule of thumbs - 30 for continuous data, 50 for discrete data.) | **Step 6:**  Then click on the “**Sample**” button. (This now takes a simple random sample of 200 males and 200 females from the dataset). |



## Exercise:

Go to NZGrapher and explore one of the following two datasets. Then choose a group variable, and then take a sample from each group.

1. Diamonds.csv

2. Rugby.csv

# Data

Data is either **observational** or **experimental**. 

**Observational** data can be collected through different methods.

* Observations
* Interviews
* Questionnaire
* Database

## Example:

The class investigations are all observational data.

## Cleaning data

Look for the following issues:

* Data entry mistakes
* Incorrect units
* Missing data

If you are **CERTAIN** the data is wrong, then make the cell blank (or enter a 0).



#### Exercise:

On Google Classroom, you can find the datasets for the class investigations. Choose one of the datasets and clean the data.

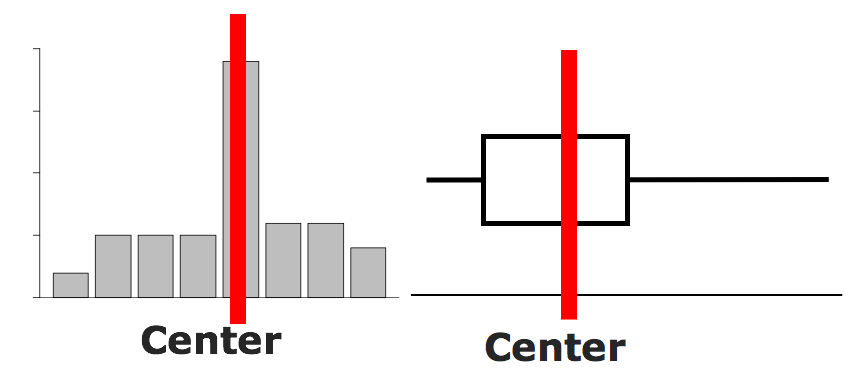
* PE Wall sit
* Social Media
* Bungee
* Sphero Racing

# Analysis

## 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



There are 3 measures of center:

* Mean =
* Median = the number in the middle (when the data is in order)
* Mode = the most common number

### Example:

Find the mean, median and mode for this data: 9, 3, 1, 8, 3, 6

Mean = = 5

Median

Put the numbers in order: 1, 3, 3, 6, 8, 9

Find the number(s) in the middle: 1, 3, 3, 6, 8, 9

Find the median = = 4.5

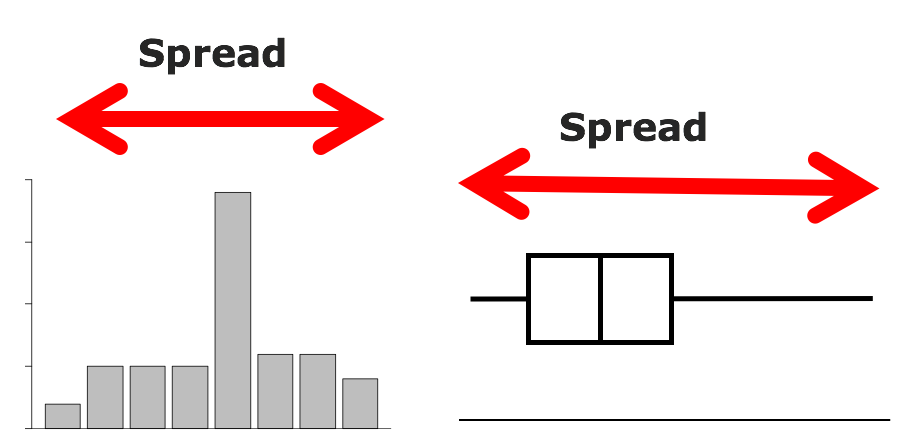
Mode = 3

### Exercises:

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

1. 4, 6, 3, 8, 2, 4, 9
2. 4.4 4.7 3.5 2.2 4.2 6.7 2.9 4.4 1.5 2.0 3.3
3. 25, 35, 37, 36, 28, 29, 36, 26, 22
4. $150, $145, $135, $150, $148, $156, $143

## Measures of Spread



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

* Range = Maximum - Minimum
* IQR (Inter Quartile Range) = UQ - LQ

where UQ = Upper Quartile = the number where one quarter of the data lies **above** it (find the median, then find the middle of the numbers **above** the median, this is the UQ),

and LQ = Lower Quartile = the number where one quarter of the data lies **below** it (find the median, then find the middle of the numbers **below** the median, this is the LQ).

### Example:

Find the range and IQR for this data: 9, 3, 1, 8, 3, 6

Range = 9 - 1 = 8

IQR

Put the data in order: 1, 3, 3, 6, 8, 9

Find where the median is: 1, 3, 3 | 6, 8, 9

Find the LQ (the median of numbers below the median), the median of 1, 3, 3

LQ = 3

Find the UQ (the median of numbers above the median), the median of 6, 8, 9

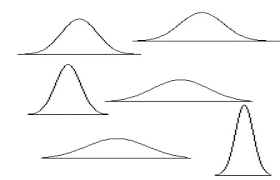
UQ = 8

IQR = UQ - LQ = 8 - 3 = 5

### Exercises:

For each of the following data sets, calculate the Range, and the Inter-Quartile Range

1. 35, 48, 36, 24, 19, 56, 43, 23
2. 3.5, 4.2, 2.6, 3.9, 2.8, 3.9, 4.2
3. $45, $35, $56, $29, $89, $76, $83, $74, $21, $42
4. Circle the graph which has the largest spread:



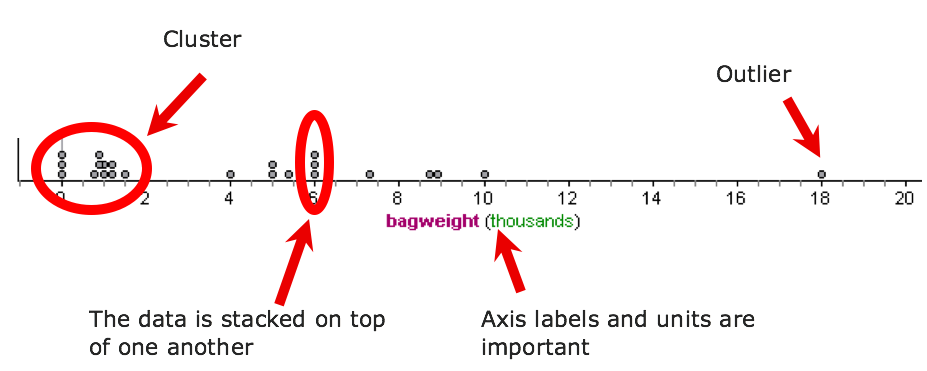
## Displaying data

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.



### Exercise:

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

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

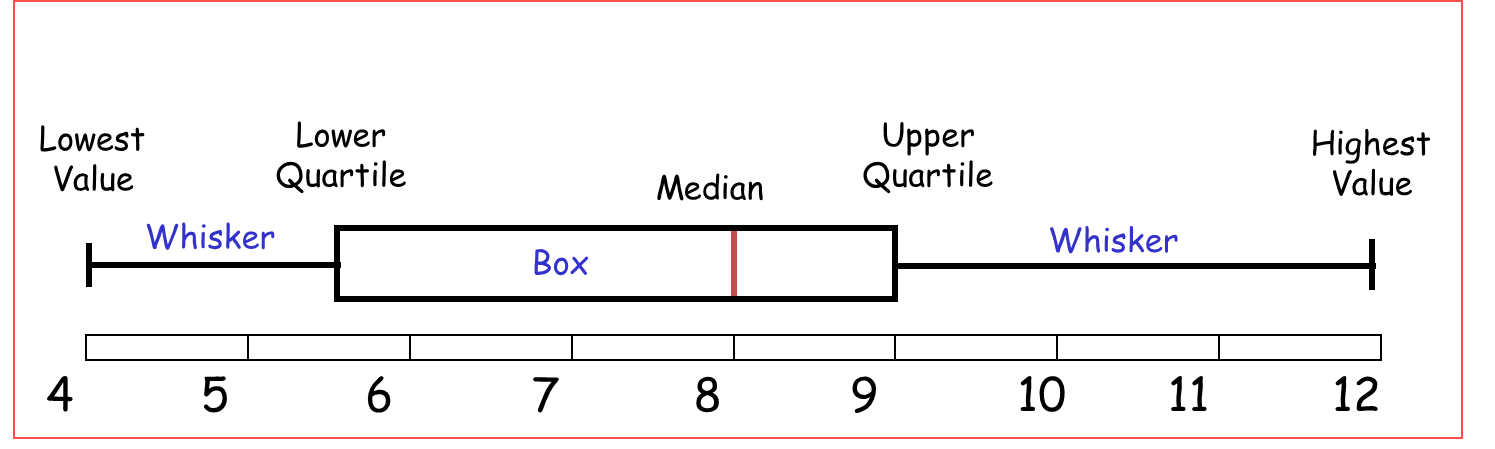
Draw an axis with the appropriate scale and label, and plot the points to create a dot plot of weekly income.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

## The 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.



### 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.

Minimum = $456

LQ = $504

Median = $768

UQ = $1,248

Maximum = $1,464

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |



Now that you know how to calculate summary statistics and draw the two types of graphs (dot plot and box and whisker), the good news is that we will now get the computer to do the calculations for us.

We need to know how to interpret them, and what they mean.

### Exercise:

Go to NZGrapher and create comparative dot plots, box plots and summary statistics for one of the following 2 datasets:

1. Diamonds.csv

2. Rugby.csv

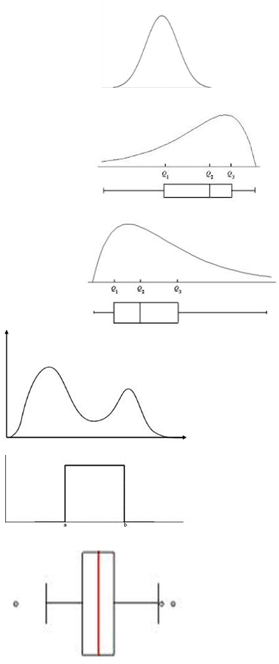
## Writing Comparative Statements

Here are the features you need to analyse and **COMPARE**.

1. Shape
2. Center
3. Spread

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

## 1. Shape Exercise

**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)

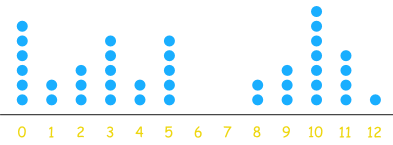
**Bimodal**

(there are two peaks)

**Uniform**

(the sides are straight and it looks like a box)

### Example:

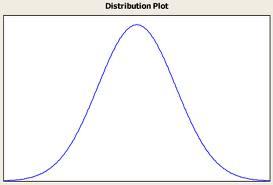
 

The data has 2 peaks, so looks approximately bimodal in shape.

### Exercise:

Sketch over the top of each graph and then state what shape it most closely matches.

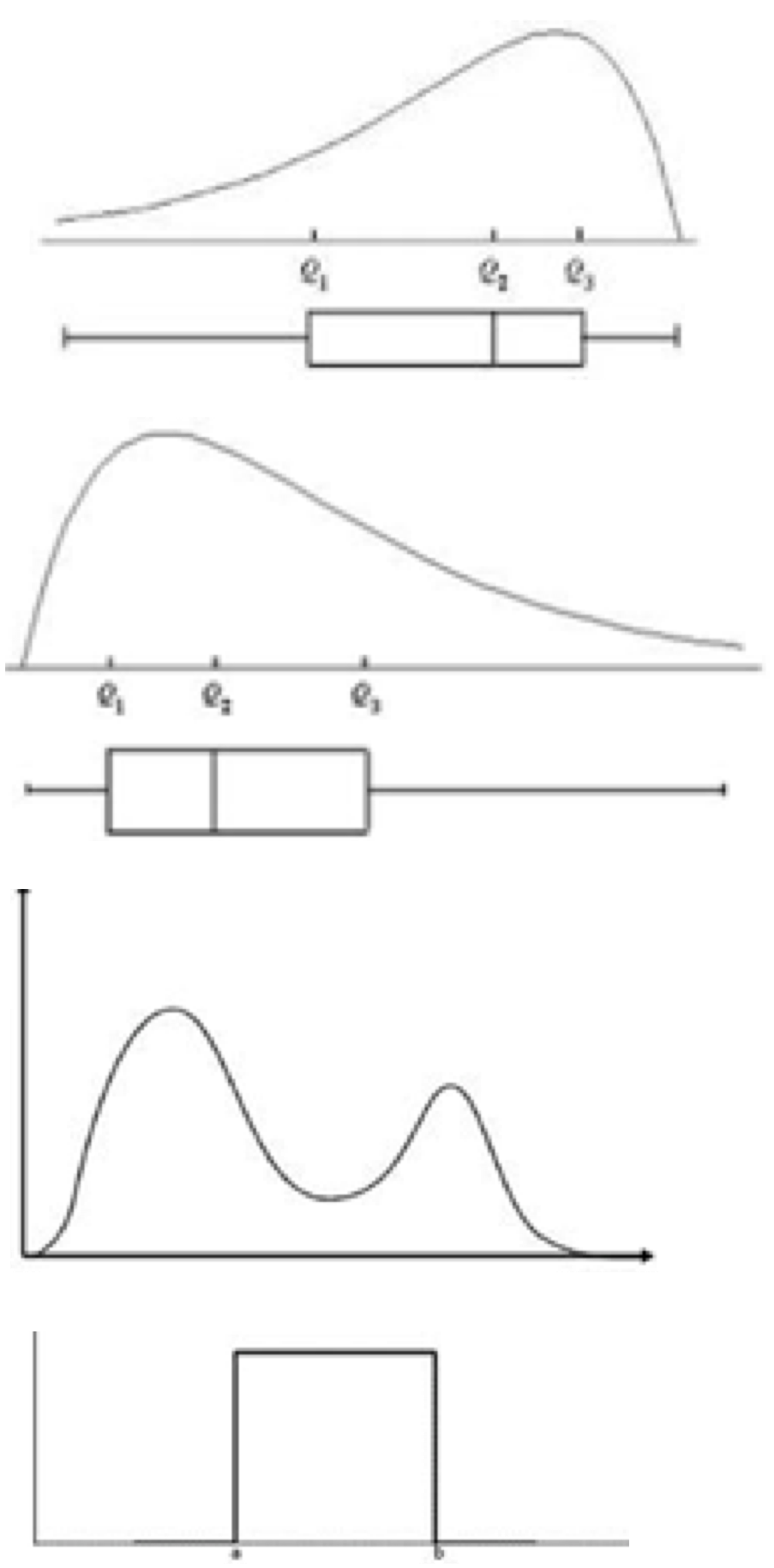
|  |  |  |
| --- | --- | --- |
| 1. | 2. | 3. |
|  |  |  |
| 4. | 5. | 6. |
|  |  |  |
| 7. | 8. | 9. |
|  |  |  |
| 10. | 11. | 12. |
|  |  |  |
| 13. | 14. | 15. |
|  |  |  |

**Justifying Shape Example (Merit)**

I notice that my data follows an approximate Normal distribution shape. This is because it is a mound shaped, symmetrical curve.

I am not going to write an example of each shape for you, as it is really important for that you start developing your own style and descriptive words.

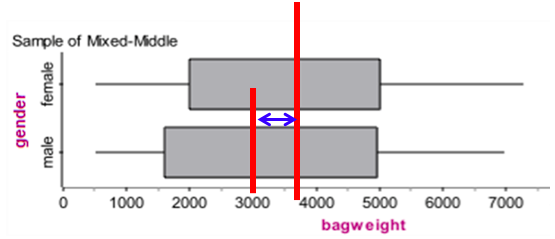
**Exercise**

For each graph, describe and justify the choice of shape.

## Comparing the Centers

Locate the medians, and tell me whichgroups **median** is **bigger** and by **how much**.

For Merit, you need to add the justification and evidence.



## Spread – comparing the spread

Find the IQR, and tell me which group’s **spread** is **bigger** and by **how much**.

For Merit, you need to add the justification and evidence.



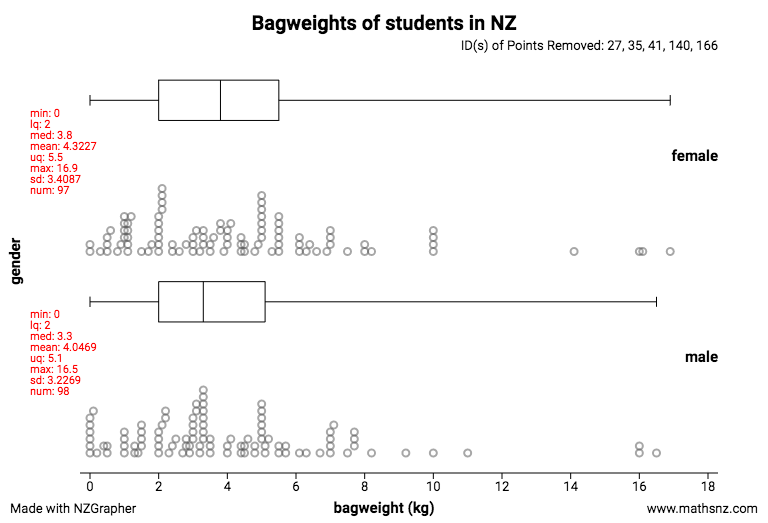
### 

### 

### Example:

**Problem:**

Does the weight of girls’ school bags tend to be greater than the weight of boys’ school bags in the 14 middle schools in NZ?



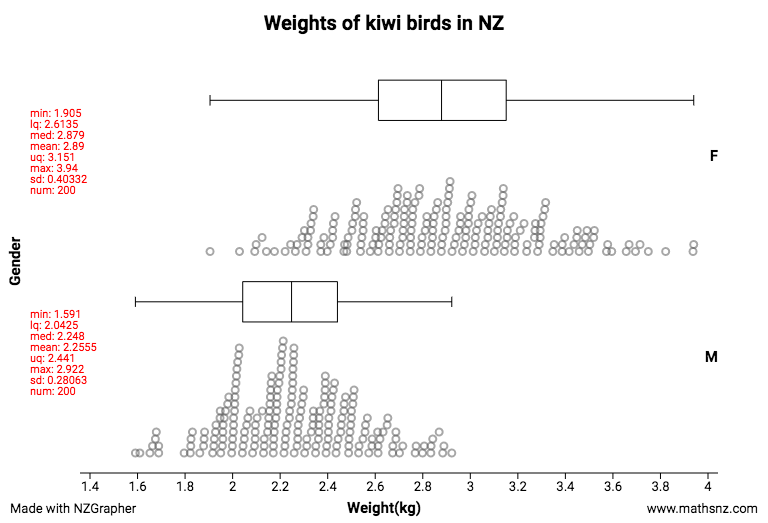
For my sample, I notice:

* The shape of the females and male bag weights in the 14 middle schools have the same right skewed shape. The females and male bag weights are right skewed because the peak is on the left hand side and there is a longer tail on the right hand side.
* The median of the female bag weights in the 14 middle schools appears to be a little heavier than the bag weights for males by 0.5kg. My evidence is that the median bag weight for females is around 3.8kg while the median bag weight for males is around 3.3kg.
* The spread of the middle 50% of females bag weights is slightly larger than the spread of males bag weights, because the IQR of the females is approximately 3.5 kg compared to the IQR for males of 3.1 kg.

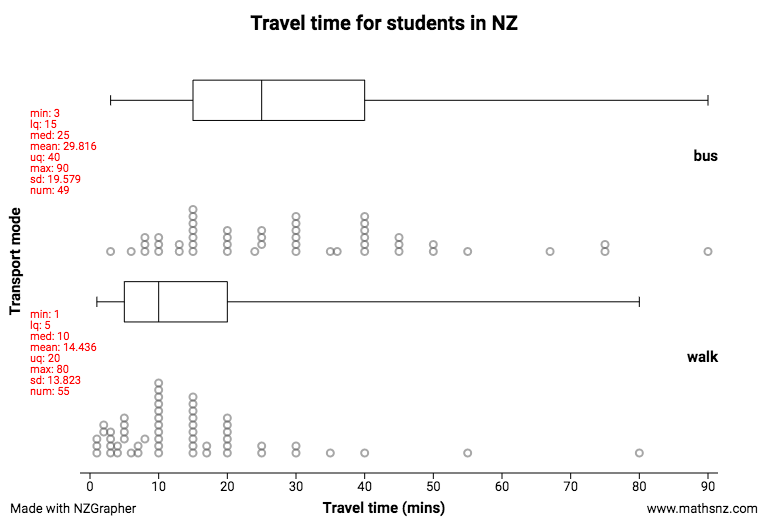
### Exercise

For the samples below, write a complete analysis. Discuss features such as: shape, center and spread.

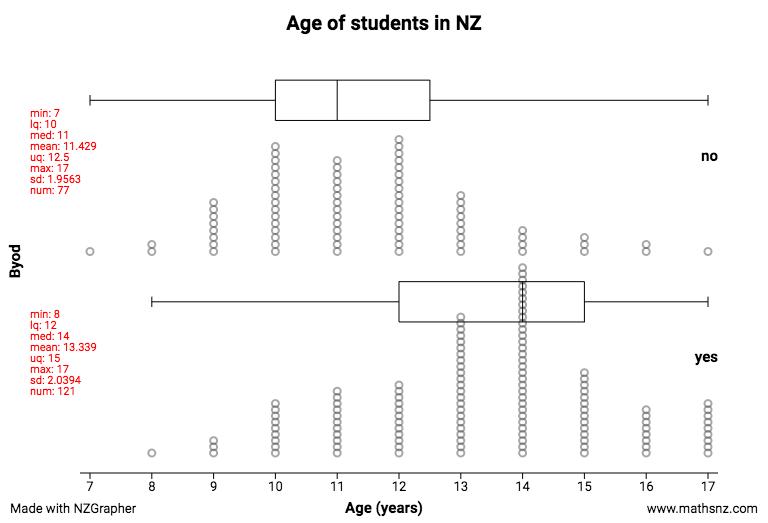
1. I wonder if male kiwi birds tend to be heavier (kg) than female kiwis, from kiwis across NZ?



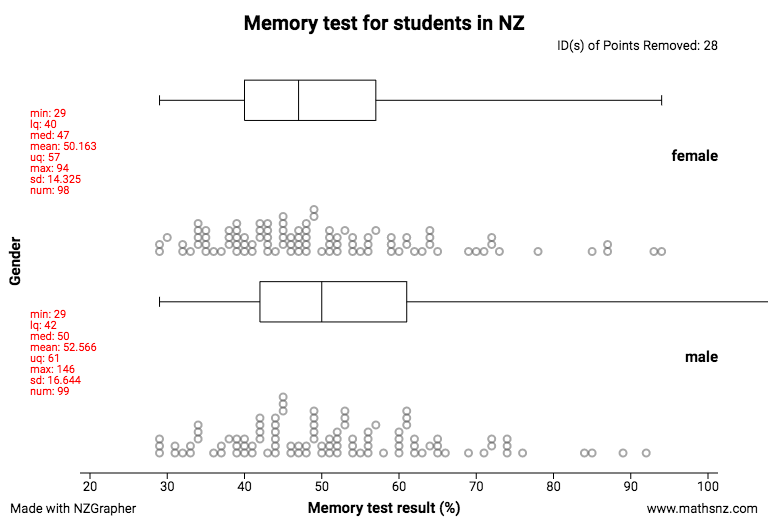
1. From high school students in New Zealand, do students who take the bus to school tend to have a longer travel time (minutes) than those who walk to school?



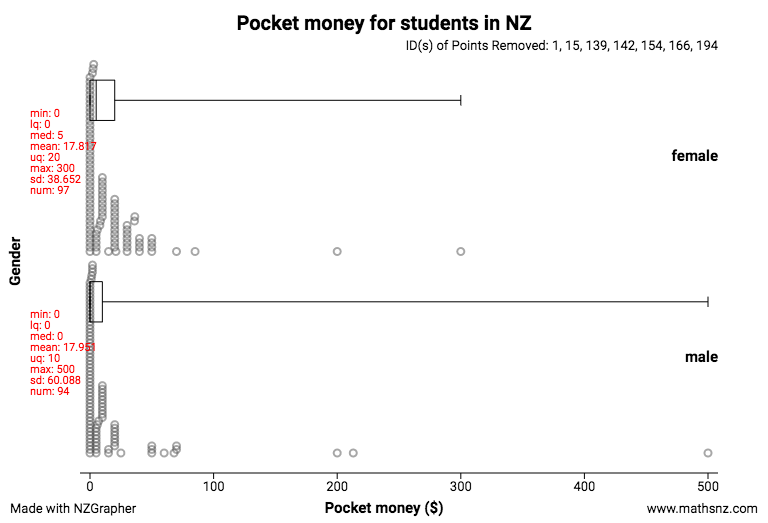
1. From high school students in New Zealand, do students who have a device tend to be older (age) than those who do not have a device?



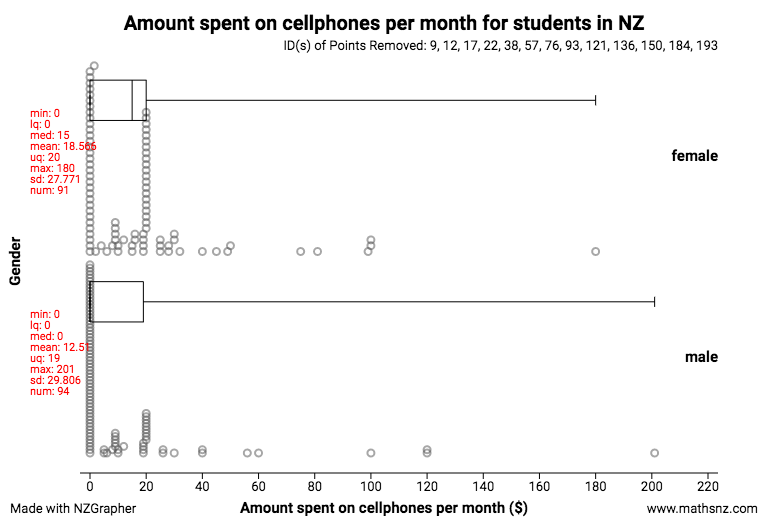
4. I wonder if female high school students from schools in NZ tend to have a higher memory test result (%) than male high school students.



5. I wonder if female high school students from schools in NZ tend to get more pocket money than male high school students.



6. I wonder if female high school students from schools in NZ tend to spend more money per month on their cellphones than male high school students.





# Conclusion

## Making the call

You need to decide if you have **enough evidence** to say if there is a **large enough** difference between the **medians** of 2 groups.



## Evidence

Locate the medians.

Do either of these lie outside the box of the other group?



### Example

Can we make the call to say that the median bag weight (kg) of males and females are far enough apart to be significant?

Or in other words, is the median weight of bags (kg) of males significantly different to the median weight of females bags, for high school students in New Zealand.

Check the median of the females. Does it lie inside or outside the middle 50% box of males’ bag weights?

Then check for the median of the males. Does it lie inside or outside the middle 50% of females’ bag weights?

**Conclusion**

**Making the call and answering the investigation question:**

I can’t make the call, because the median of males bag weight lies inside the box of the females, and the median of the females bag weight lies inside the box of the males.

**Evidence:**

Therefore for my sample, we **don’t have enough evidence** that male students tend to have heavier (grams) school bags than females do, for students from Census at School, NZ.

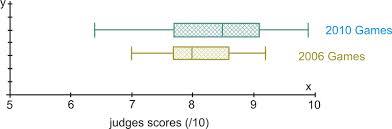
### 

### 

### Exercise:

Can you make the call for each of these boxplots? What does that mean? Is the difference between the medians large enough? Explain.

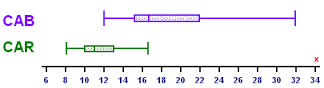
1. Judges score in Gymnastics at the Commonwealth Games.



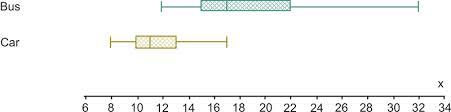
I can / can’t make the call, because

Therefore I have / haven’t got enough evidence that the judges score for the 2010 Commonwealth Games in Gymnastics tends to be more / less than the judges score for the 2006 Games.

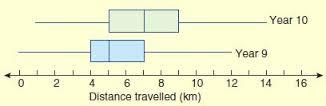
1. Cost of taking a cab or driving a car to the airport in Auckland



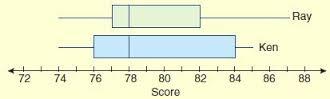
1. Cost of going from Manukau to the Airport.



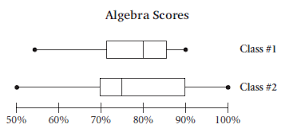
1. Distance that students in Auckland travel to get to school.



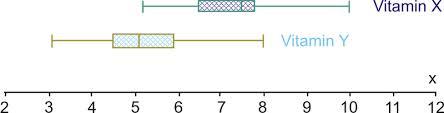
1. Points scored in cricket matches.



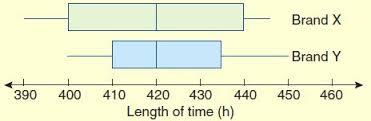
1. Marks in Algebra tests.



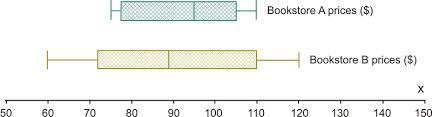
1. Vitamin levels in fruit.



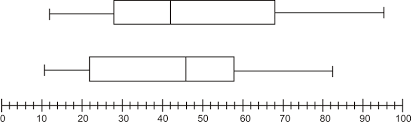
1. Time for concrete to set.



1. Prices of textbooks.



1. Time for students to eat a peanut butter sandwich (seconds).

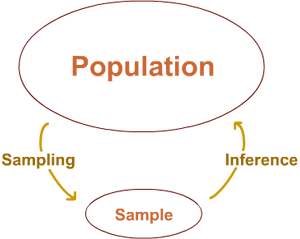


**Inference**

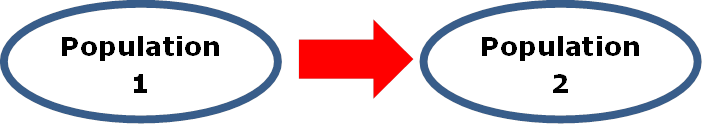
When we answer the investigation question, we are making an **inference**.

We have taken a sample, and decided whether or not you can make the call.

Now you need to answer your investigation question, which uses the information from the **sample** to make an **inference** about the **population**.



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



The conclusion can only be applied to an identical population of the one for which the data was collected.

### Example:

If the data is of Americans, then the conclusions can only be applied to other Americans.

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

### Exercise:

1. Complete the following sentence.

The population of our school includes (be specific, eg. Gender)…

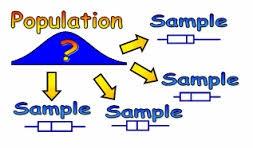
1. If we took a sample of heights from 30 students at Ormiston Senior College, would this sample be representative of the senior high school students in NZ?

Explain why/why not.

1. If we took a sample of heights from 30 students at Ormiston Senior College, would this sample be representative of the population of NZ?

Explain why/why not.

Sampling Variability

This describes the variation (differences) that happens when you take another sample.

You need to consider 3 things:

* how would the **data** change?
* how would the **analysis** change?
* how would your **conclusion** change?

Watch this animation: <http://tiny.cc/jdixpy>

It shows how the box plot changes when you take different random samples from the same population.

The graph below shows a snapshot of the animation, showing how much the variations change between sample sizes. This is the animation to watch: <http://tiny.cc/qkixpy>



You should notice that smaller sample sizes show more variation, while larger sample sizes have less variation.

### 

### 

### Exercise:

1. Do you want the sampling variability to be large or small? Why? Explain.
2. If you have two samples of students and have measured their heights, would the data (e.g. the heights of students) be **EXACTLY** the same for both groups? Explain.
3. For the two samples, compare the median heights. Would they be identical, similar, different …? Explain.
4. For the two samples, is the conclusion likely to be the same or different? Explain.
5. Complete the sentence:

If I took another sample ….

### Example

**Problem:**

I wonder if the length of the right foot for high school boys tends to be longer than the high school girls for students across NZ high schools.

**Conclusion:**

**Making the call and answering the investigation question:**

I cannot make the call that girls tend to have longer right foot length than boys, for my sample of students at high schools in NZ.

**Evidence:**

The evidence I used to make this call is that the median length of the girls lies inside the box of the boys ring finger length, and vice versa.

**Inference:**

My sample is a random sample, and so the data is likely to be representative of all high school students in NZ. This means that my conclusion for my sample is likely to be applied to high school students in NZ. I infer that that I don’t have enough evidence to support boys tending to have longer right foot lengths than girls for **ALL** students across NZ high schools.

**Sampling variability:**

If I took another sample of different girls and boys at high schools in NZ, then I would get different measurements of lengths of right foot as I would be measuring different students.

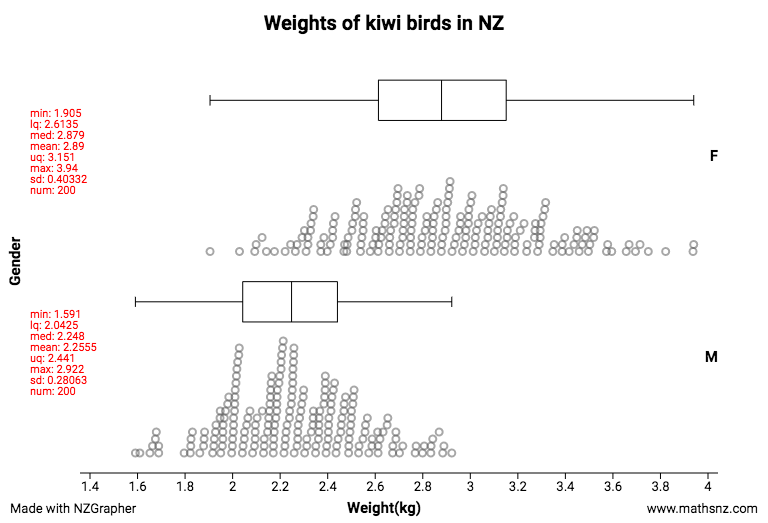
With a different sample of students, the median length of right foot from girls and boys is likely to be similar, but not the same as I found in my investigation here.

With a different sample, I would however expect that the conclusion I came to, that boys do not tend to have longer right foot lengths than girls for **ALL** students across NZ high schools, to be the same.

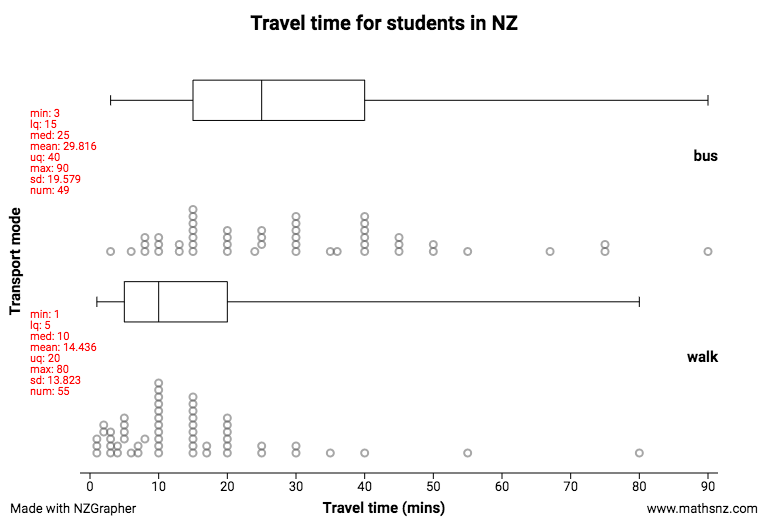
### Exercise:

For the samples below, write a conclusion. Decide whether or not you can make the call, answer the investigation question, and discuss sampling variability.

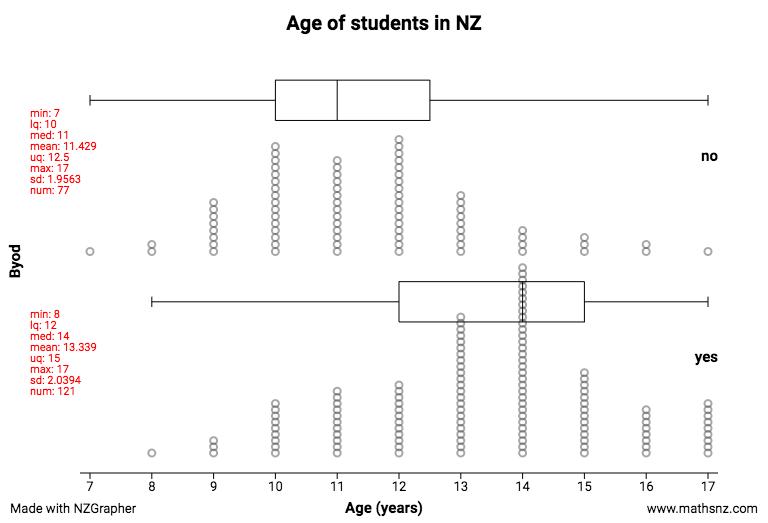
1. I wonder if male kiwi birds tend to be heavier (kg) than female kiwis, from kiwis across NZ?



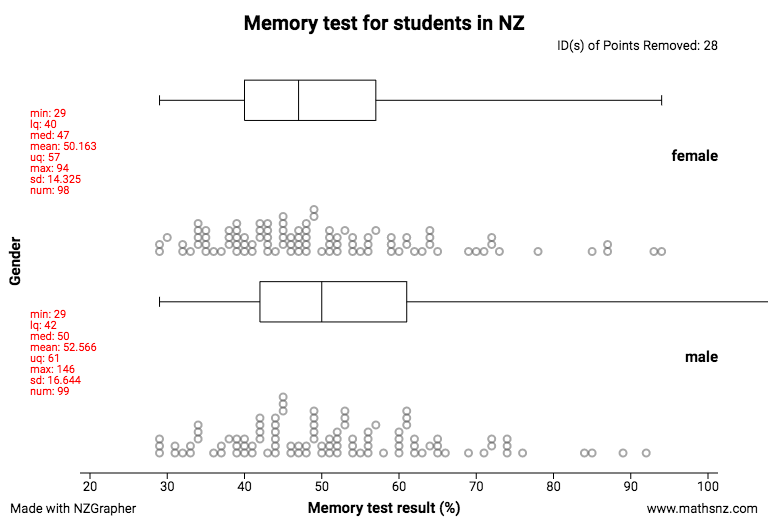
1. From high school students in New Zealand, do students who take the bus to school tend to have a longer travel time (minutes) than those who walk to school?



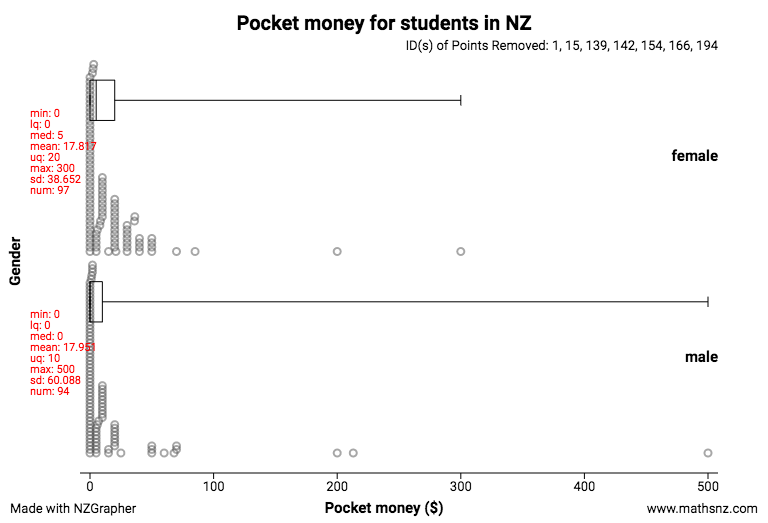
1. From high school students in New Zealand, do students who have a device tend to be older (age) than those who do not have a device?



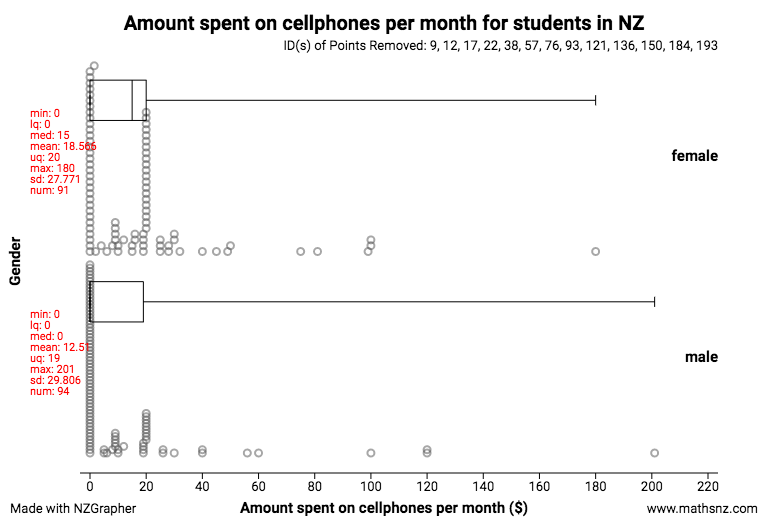
4. I wonder if female high school students from schools in NZ tend to have a higher memory test result (%) than male high school students.



5. I wonder if female high school students from schools in NZ tend to get more pocket money than male high school students.



6. I wonder if female high school students from schools in NZ tend to spend more money per month on their cellphones than male high school students.





## 

Fill in the planned completion dates, and when you have finished each Key skill, show your teacher so they can track your progress.

|  |  |  |  |
| --- | --- | --- | --- |
| **Key Skill** | **Workbook pages** | **Planned Completion Date** | **Completion Date (teachers sign)** |
| Problem | 1 - 5 |  |  |
| Plan / Data | 6 - 7 |  |  |
| Analysis | 8 - 25 |  |  |
| Conclusion | 26 - 43 |  |  |