Level 2 – AS91264

Standard 2.9

4 credits – Internal

Use statistical methods to make an inference

Contents

Lesson Zero: Revision of 1.10 multivariate data

PPDAC

Last year in 1.10 multivariate data, you were given a sample and had to investigate it using the PPDAC cycle. It stands for: Problem, Plan, Data, Analysis, and Conclusion.

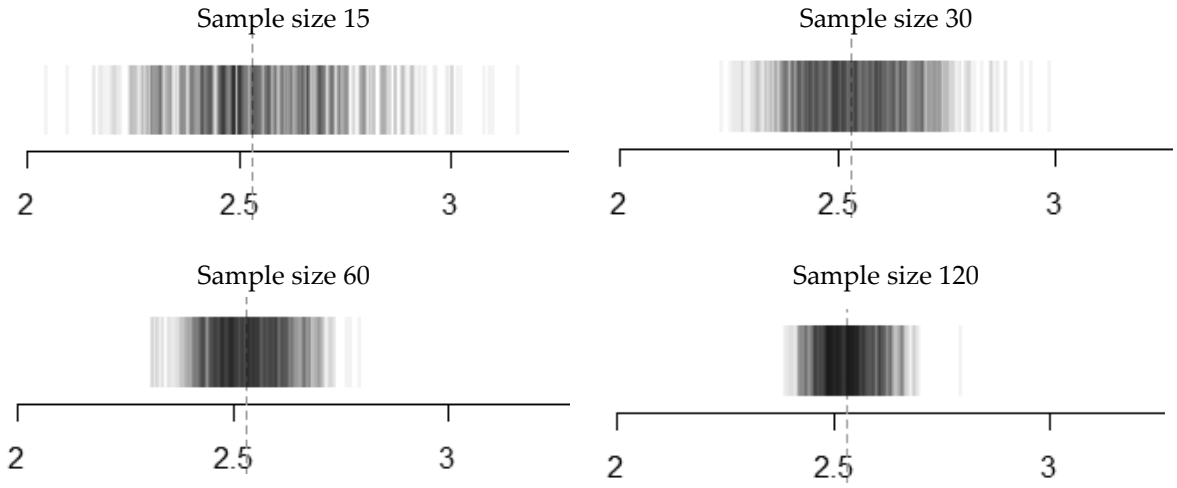
Sampling

You were wanting to make a claim about the entire population, e.g. "Do year 11s' schools bags tend to be heavier than year 10s' school bags?", but you only had access to a sample of the population. You didn't have data for every year 10 and year 11 student in New Zealand.

Samples, unlike censuses, have sampling variation.

Sampling variation

Sampling variation causes different samples to have different results even though they're drawn from the same population. You can minimise sampling variation by increasing the sample size but you can't eliminate it without taking a census. Sampling variation also depends on the spread of the population data. Below are representations of sampling variation for different sample sizes.



Sampling variation cont.

When you graph your data in a box and whisker plot, there is this invisible 'fuzziness' around them caused by sampling variability.

Whenever you see…

You remember…

Lesson Zero questions

Q1. Define 'population'.

Q2. Write up to a paragraph explaining sampling variation.

Q3. What effect does sample size have on sampling variation?

Q4. What is the only type of investigation that doesn't have any sampling variation?

Q5. When you see:

You remember:

Lesson One: Investigative question (Problem)

Introduction to comparing medians
Last year, in 1.10 multivariate data, you asked a question along the lines of:

"I wonder if the school bags of year 11 students are heavier than the school bags of year 10 students in New Zealand?"

This year, you'll be comparing medians. For example:

"I wonder if the median weight of school bags of year 11 students is heavier than the median weight of school bags of year 10 students in New Zealand?"

In other words, you'll be making an inference about the population medians from the sample medians.

What is the median and how do you calculate it?

Crafting an investigative question
Crafting your investigative question requires you to state:

- The quantitative variable that is being investigated, e.g. bag weight
- The groups being compared. This is also known as the qualitative variable, e.g. year 10s and year 11s
- The population from which you are sampling, e.g. New Zealand students
- The direction of the comparison (e.g. "Is the median height for year 11 girls *taller* than…")

It also pays to include a few extra details to give your report more impact:

- Context about the situation (e.g. "Adolescence is a time of great physical growth but different genders reach puberty at different times")
- A reason for the investigation (e.g. "Students often carry very heavy school bags. Is this causing them any harm? I will investigate if the median weight of year 11s' bags is heavier than…"
- Your hypothesis. What do you think it will be? Why do you think that?

For example, "Adolescence is a time of great physical growth. The common belief is that boys are taller than girls. But girls reach adolescence before boys so they could be taller, on average, for a time. Or perhaps the belief that boys are taller than girls is a myth. I will investigate whether the median height of year 12 girls is taller than the median height of year 12 boys in New Zealand. My hypothesis is that the median height of boys will be taller than girls because, based on the year 12s that I know, boys tend to be taller."

Lesson Two: Representative sampling (Plan)

Sampling

You are required to take a sample to make your inference with. Any sample-to-population inference we make is only as good as your sample. There are two things to consider when taking a sample: sample size and sampling method.

Sample size

You have already seen how a larger sample decreases sampling variation. But there are several factors that limit the maximum sample size you can reasonably take:

- It may take too long. Not only does a long survey cost most money (or volunteer time) but conditions can also change during the survey. For example, if you were surveying people regarding their opinions of the government's response to boy racers and a major crash involving a boy racer occured during the survey, the opinions of the population will likely change.
- The sampling process may be destructive. For example, if you are testing packaged food for their ingredients or electrical components for their life, the testing process destroys the product. Clearly, you wouldn't want to test more than you need to.
- Not every element of the population may be available for investigation. For example, if you were wanting to investigate the current population of Blue Cod, an endagered fish endemic to New Zealand, it is impossible to find all of them. Your sample would be limited to those that you catch.

Valid sampling methods

A valid sampling method will give you a representative sample of the population. The valid sampling methods are:

Random sample

Obtaining a random sample is difficult but gives a truly random sample which is representative of the population. It is difficult because it requires that all of the elements in the population can be numbered.

To take a random sample, number all of the elements in the population from 1 to the total number of elements. Generate random numbers between 1 and the total number to select elements to be a part of the sample (usually by using a calculator or a computer). Any repeats are ignored. Keep generating random numbers until you have the desired sample size.

Systematic sample

If you can't number all of the elements in the population, you may be able to take a systematic sample. If you can order the population in some way then you can take a systematic sample. For example, alphabetical order, seating order, the order along a production line, etc.

To take a systematic sample, order the population if it's not already ordered. Then, select every nth element, e.g. selecting every 13th biscuit along a production line. Make sure that there is no natural pattern within the ordering of the population, e.g. if every 10th biscuit happens to be in the oven for longer than the rest, then selecting every 10th biscuit isn't representative of the population.

Cluster sample
A cluster sample selects a group (or a "cluster") to represent the entire population. Ideally, the cluster is representative of the population. It is easy to get a cluster sample but the group will be together for some reason. That reason may make them unrepresentative of the population, depending on what variable is being investigated. For example, students at Wā Ora will be representative of all high school students if the variable being investiaged is height. But they won't be representative if the variable being investigated is the type of school they attend.

Stratified sample
If there are minority groups in the population that you want to make sure are represented, you could take a stratified sample. In a stratified sample, the proportion of groups (or "strata") in the sample is representative of the proportion of groups in the population. For example, if the population is 70% Pākehā, 15% Māori, 10% Asian, and 10% Pacific Islander then, in a stratified sample, the proportions would be similar or the same.[1] You could stratify by age, gender, employment status, or any other variable you can think of.

Invalid sampling methods
There are other sampling methods that give you an unrepresentative sample which are not valid. They are:

Self-selected sample
If the respondent does something to be selected then it is a self-selected sample. These are invalid as those that respond tend to have strong opinions. Examples include: online surveys and text-to-vote polls.

Person in the street
Reporters will often approach people in the street to ask their opinions on a news story. There is little credibility in this method as the interviewer may only approach people they think have something to contribute (or are more approachable). Furthermore, all of the people are in the same place at the same time which may make the sample less representive of the population.

Quota sampling
In quota sampling, the researcher is given a quota of minority groups to include in the survey. While this sounds similar to stratified sampling, it has been shown to be invalid as the researcher can tend to choose individuals who meet their quotas as quickly as possible, for example an individual who is unemployed, disabled, female, and foreign.

When you're ready, attempt the following questions.

[1] The ethnicities don't add to 100% because some people identify with more than one ethnicity.

Lesson two questions

For the following questions, comment on the source of bias which makes the sample unrepresentative.

Q1. A newspaper wnats to know the opinions of the people in Napier on amalgamation with Hastings. They conduct a street interview outside Napier Countdown between 3pm and 5pm on Wednesday.

Q2. A newspaper conducts a phone survey of 500 people in Wellington on twhether they are going to vote for New Zealand first, and applies the results to all of New Zealand.

Q3. The school principal wants to know how much homework is being done by year 9 students, so she visits one year 9 class and asks the students in that class.

Q4. The board of trustees wants to know whether or not the students want to change the school uniform so they give a quiestionnaire to the first five students on the roll of each class.

Q5. The local council wants to know if people in the area want more or less council-owned parks. They send out a survey to every household with three or more people in it.

Q6. A lobbying group wants to know the public's views on abortion so they stand outside the hospital, asking people for their views.

Lesson Three: Analysis

Analysing your sample
Before you make your inference, you are required to familiarise yourself with the sample. This is also called analysis. Your analysis should include comments on the:

- Centre and/or middle 50% of the data
- Shift or overlap
- Shape
- Spread
- Unusual features (including grouping)

Often, these comments will flow back and forth between each other. They are not necessarily discrete and unrelated concepts.

Centre and middle 50%
This regards the typical values of the data. You could consider the center (a single value) or the middle 50% (a range of values).

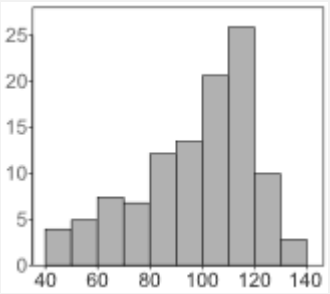
The center can be described by either the median or the mean. They are similar but different. The median is the middle value whereas the mean is the sum of the values divided by the number of values. The mean is affected by nonsymmetrical data and extreme values.

The middle 50% is described by the quartiles. It is the range between the upper and lower quartiles.

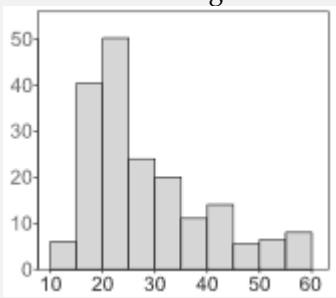
Shift and overlap
Shift describes the extent to which the data is "shifted" from one group to the other. The overlap describes the extent to which the data overlaps between the two groups. They are opposite measures; a large shift is the same as saying a small overlap and vice versa.

Shape
The data could be symmetrical with a single peak (unimodal). Or it could have more than one peak (bimodal, trimodal, etc.). Or it could be nonsymmetrical (skewed to the left or skewed to the right).

Skewed left



Skewed right

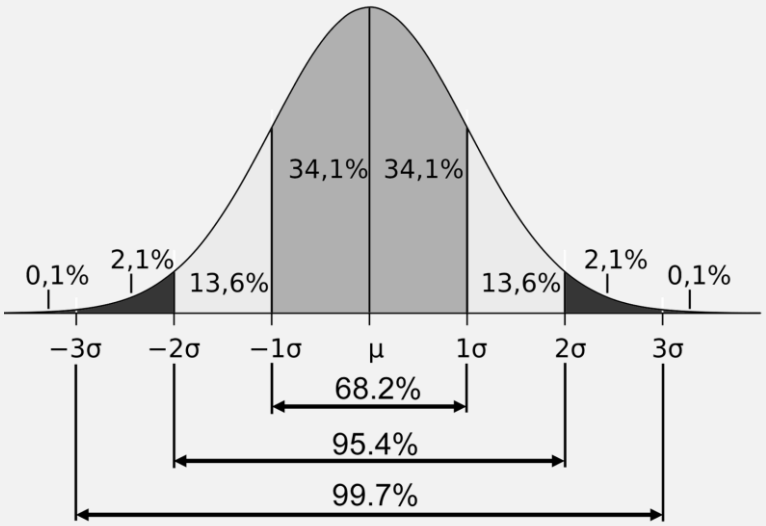


Spread

The spread of the data is how wide it is spread from the centre. The spread can be seen in the graph and backed up by either the interquartile range (IQR; the distance between the upper quartile and the lower quartile) or the standard deviation.

Because the IQR is based on the quartiles, it is in turn based on the median. Whereas the standard deviation is based on the mean. The standard deviation is a powerful measure of spread. Assuming the data is normall distributed (symmetrical and unimodal), the following is true:

- 68% of the data is within 1 standard deviation from the mean
- 95% of the data is within 2 standard deviations from the mean
- 99.7% of the data is within 3 standard deviations from the mean



If the standard deviation is low, the spread is low and vice versa.

Unusual features

If there is anything else that's noteworthy about your data, then you should describe it as an unusual or interesting feature. You could mention if there are groups (or clumps) within the data.

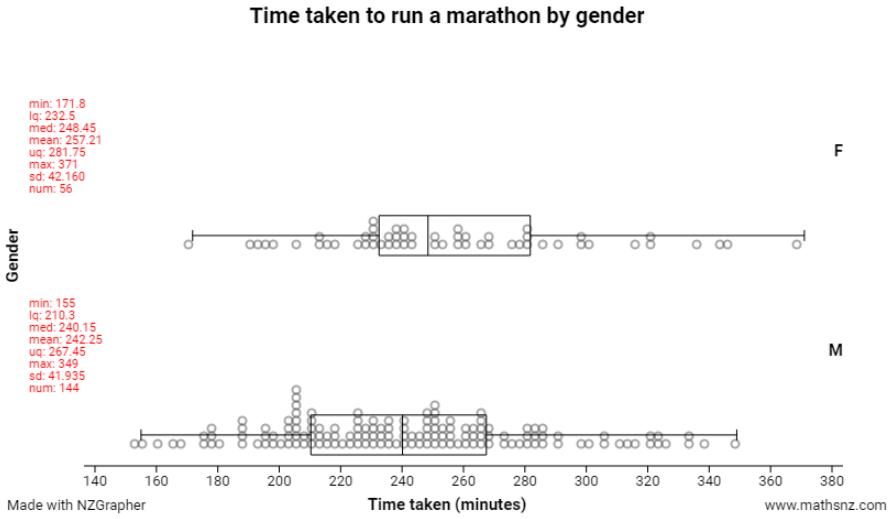
Signal vs. Noise

For all of your comments in your analysis, try to think about signal vs. noise, i.e. is the difference or notable feature you're seeing probably due to an actual difference or feature in the population (signal) or sampling variation (noise)?

Lesson Four: Inference

Reintroduction to inference
You already have a sample for both of the groups you are comparing. Each group has its own summary statistics, including median. You can therefore say (with certainty) which group has the fastest/heaviest/biggest, etc. median *in your sample*.

Consider the data below:



Answer the question: "For this sample, which gender has the fastest median time to run a marathon?" Include the median time for each gender in your answer.

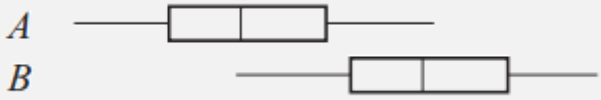
But your investigative question is about the population, not just the sample. Do you remember this?

Explain sampling variability.

Inference in 1.10 multivariate data

In 1.10 multivariate data, you made an inference based on the following rules:

If one lower quartile is bigger than the other upper quartile

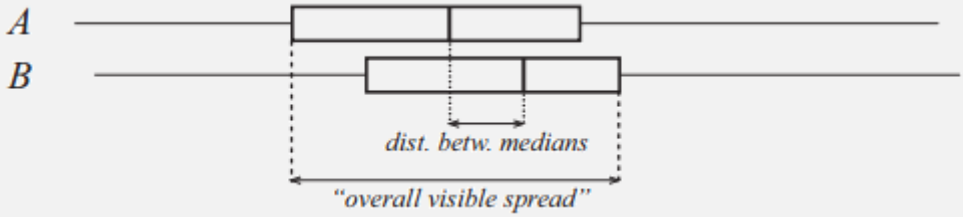


Then you can say that the median for B will be larger than the median for A in the population.

If one median is outside the other middle 50%

Then you can say that the median for B will be larger than the median for A in the population.

Comparing the distance between the medians (DBM) to the overall visual spread (OVS)



You can say that the median for B will be larger than the median for A in the population if:

- The DBM is greater than about a third of the OVS for sample sizes around 30
- The DBM is greater than about a fifth of the OVS for sample sizes around 100
- The DBM is greater than about a tenth of the OVS for sample sizes around 1 000

Inference in 2.9 Inference

Now you will use a more sophisticated tool to allow you to make more accurate inferences. You want to be able to give an interval in which you're "pretty sure" the population median is. If there is a large spread in the data, this interval will be wider. If there is a large sample size, this interval will be smaller. It is called the informal confidence interval. It has the equation:

$$\text{Informal confidence interval} = \text{Sample median} \pm 1.5 \times \frac{IQR}{\sqrt{n}}$$

where IQR is the interquartile range and n is the sample size.

In other words, the informal confidence interval is the sample median plus or minus 1.5 times the interquartile range, divided by the square root of n, the sample size.

This gives a range in which you can be "pretty sure" the population median is in. In fact, this is a 95% informal confidence interval, meaning you can be 95% sure that the population median is within the interval.

Lesson four questions

For the following questions, use the given information to calculate the 95% informal confidence intervals.

Q1.

Sample median	10.6kg
IQR	5.8kg
Sample size	30

Q2.

Sample median	1.5m
IQR	0.3m
Sample size	36

Q3.

Sample median	161 min
IQR	24 min
Sample size	68

Q4.

Sample median	2.55kg
IQR	741g
Sample size	35

Q5.

Sample median	$281
LQ	$268
UQ	$302
Sample size	52

Q6.

Sample median	40 years
LQ	35 years
UQ	47 years
Sample size	121

Q7. What happens to the 95% informal confidence interval as the spread increases?

Q8. What happens to the 95% informal confidence interval as the sample size increases?

Q9. How big does the sample size need to be to minimise the spread?

Lesson Five: Conclusion

Answering the investigative question
Remember what your investigative question is. It is something along the lines of "Is the median [variable] for group A bigger than the median [variable] for group B in [the population]?" For example: *"Is the median height for year 11 girls taller than the median height for year 11 boys in New Zealand?"*

The conclusion is where you answer that question based on your 95% informal confidence intervals. You must interpret both of the intervals, e.g. "I can be pretty sure that the median height for year 12 boys in New Zealand in between 161cm and 185cm."

If the 95% informal confidence intervals have no overlap, that means that you can be "pretty sure" that the difference (or "shift") you see in your sample reflects a real difference back in the population.

If the 95% informal confidence intervals have *any* overlap, that means that there is a reasonable chance that the sample medians are actually the other way around in the population so you can't answer either yes or no. In other words, the sampling variability is obscuring the population too much for you to make a call about which group's median is larger.

Showing an understanding of sampling variability
Another thing you must do in your conclusion (if you haven't already in your analysis) is to show your understanding of sampling variability, i.e. that samples vary from one another.

Note well: Even though the sample medians will change from sample to sample, you expect that the inference remains the same (e.g. that either one group's population median is larger or that you can't say). That's the whole point of inference.

For example: *"If I had another sample from the sample population, I expect that there would be some variation in the centre, shape, and spread of the data. The sample medians and 95% ICIs would likely be different too. However, I would still be 95% confident that the different ICIs contain the population median so I would expect to make the same inference".*

Additionally, the conclusion is also the place to discuss the impact of one aspect such as sample size if you haven't already. Basically, the larger the sample, the less sampling variation which means that the sample centre, shape, and spread more closely match the population features. Also, the 95% ICIs will be narrower because you can make better estimates for the population medians. You may have already mentioned this in your plan or your analysis.

Finally, the conclusion is also the place to summarise the key results from your analysis and explain what they could mean for the population. For example: *"The girls data is much more spread in the sample, having an IQR of $420 compared to $120 for boys. It makes sense that because girls tends to spend more, they would also tend to spend more varying amounts."*

You could also consider the context and if there are other reasons for features in the data. For example: *"Girls' clothes and accessories vary much more in price compared to boys' clothes. I've seen dresses cost anything from $100 to $600 whereas suits only tend to cost between $80 and $150. Also, girls have more things to spend money on like their hair and jewellery which boys don't usually spend money on."*

Sample conclusions

Here are four conclusions at the four different levels of achievement to give you an idea of what is expected. The data is the same as the example analyses: amount spent on the school ball. The investigative question was "Do New Zealand year 12 and 13 girls have larger median spends than boys on the school ball?"

Not Achieved conclusion The median spent for boys was $200 and the median spend for girls was $310. Therefore, New Zealand year 12 and 13 girls have larger median spends than boys on the school ball.	**What they did well** They copied the wording from the question.	**What they didn't do well** They confused the sample medians for the population medians. They didn't use the ICIs to estimate where the population medians are.
Achieved conclusion I can conclude that New Zealand year 12 and 13 girls have larger median spends than boys on the school ball because the confidence intervals aren't overlapping. If I took another sample, I would expect to get different sample statistics.	They based their inference on the ICIs. They talked a little bit about sampling variability.	They didn't explain what the ICIs mean or what they're for. They could explain sampling variability more.
Merit conclusion I can use my sample to estimate the population medians. Based on my 95% informal confidence intervals, it is likely that the median amount spent by girls is between $270 and $350 whereas the median amount spent by boys is likely to be between $190 and $210. Therefore, I can conclude that New Zealand year 12 and 13 girls have larger median spends than boys on the school ball. If I had smaller sample sizes, I wouldn't be able to make as good estimates for the population medians and my ICIs would be wider. However, the difference is so great that I could still make an inference will smaller sample sizes.	They explain and interpret the ICIs. They discuss the effect of changing the sample size.	They should summarise key points from their analysis. They could also consider possible explanations for their findings.

Excellence conclusion

I can use my sample to estimate the population medians. Based on my 95% informal confidence intervals, it is likely that the median amount spent by girls is between $270 and $350 whereas the median amount spent by boys is likely to be between $190 and $210. Therefore, I can conclude that New Zealand year 12 and 13 girls have larger median spends than boys on the school ball. This backs up what my analysis showed. There was a large shift between the two medians, $200 for the boys and $310 for the girls. This is a huge difference.

It makes sense to me that girls tend to spend more on the ball than boys because, as I mentioned in my analysis, girls tend to care more about their appearance and more willing to spend money on it. Also, there are more things for girls to spend money on: jewellery, hair, and shoes. Whereas boys don't tend to spend money on these, or if they do, spend less than girls. This could explain why the girls' data is skewed to the right but the boys' data is symmetrical. The girls also have much more variation in their spending which makes sense. $50 is not a lot of money if you're spending $300 but it is if you're spending $100. This could explain why the girls' IQR is much wider than the boys' IQR (which also means a much wider ICI).

If I had smaller sample sizes, I wouldn't be able to make as good estimates for the population medians and my ICIs would be wider. However, the difference is so great that I could still make an inference will smaller sample sizes.

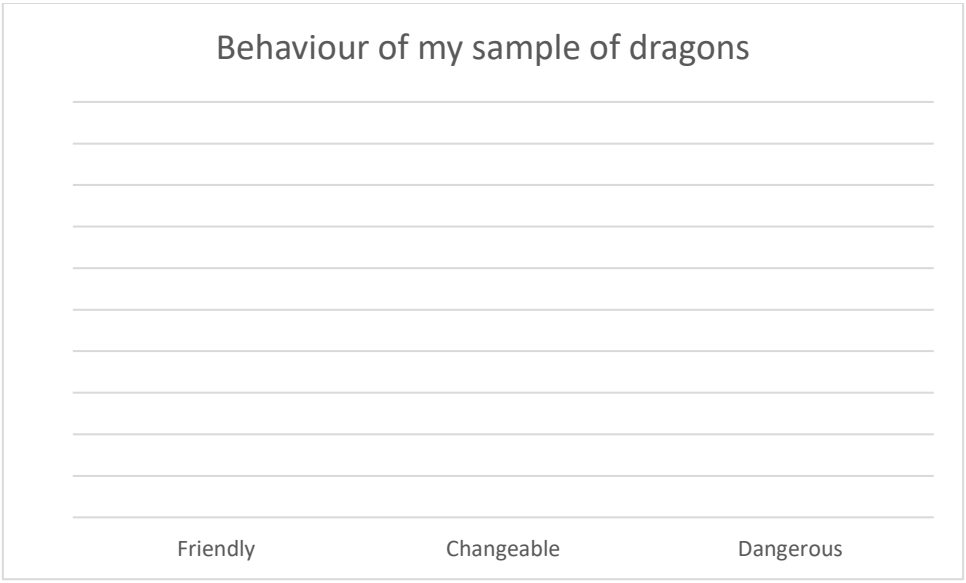
Activity One: Dragonistics data cards

Introduction to the data cards

There is a population of dragons inhabiting an island. Each dragon has a number of attributes shown below:

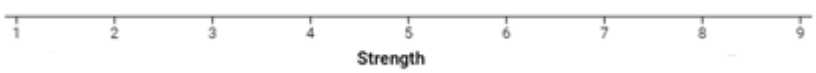
You are about to take a sample of dragons from the population. It doesn't have to be a truly random sample but make an effort to make it as representative as possible. Choose a sample size between 10 and 30 or so.

When you have your sample, note how many of each different behaviour you have and draw it on a bar graph. Calculate the proportion of your entire sample made up by dragons of each behaviour.



Now note the strength of your dragons (of all behaviours combined). Calculate the minimum, lower quartile, median, upper quartile, and maximum strengths. Show this information in a box and whisker plot. Compare this with other samples from around the class.

Minimum	
LQ	
Median	
UQ	
Maximum	



Activity Two: CensusAtSchool

Introduction to CensusAtSchool
CensusAtSchool is an online survey for New Zealand students between years 5 and 13. It gathers information on a range of variables in several categories including physical measurements, transport to school, food intake, reaction speed, and opinions.

Use the registration code to enter the survey and complete it.

Your data will be added to the database which you will eventually draw your sample from for your assessment.

Activity Three: Data

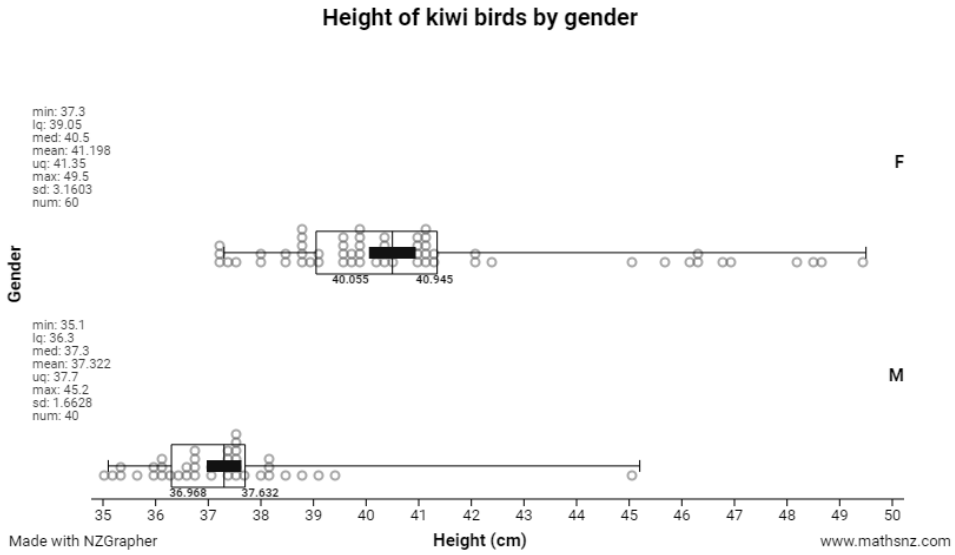
NZGrapher
NZGrapher is a free, online graphing tool designed by a maths teacher in New Zealand, Jake Wills. It can easily produce the graphs you require for this standard.

Dot plot
You are required to produce a dot plot with an appropriate title and labels.

Box and whisker plot
You are also required to produce a box and whisker plot including summary statistics. I recommend you use NZGrapher to put this on the same graph as your dot plot.

Informal confidence intervals
You haven't looked at these yet but you will need to include the informal confidence intervals and the confidence interval limits on your box and whisker plot

Produce the graphs for your assessment samples like the example below.



Activity Four: Sample median variation

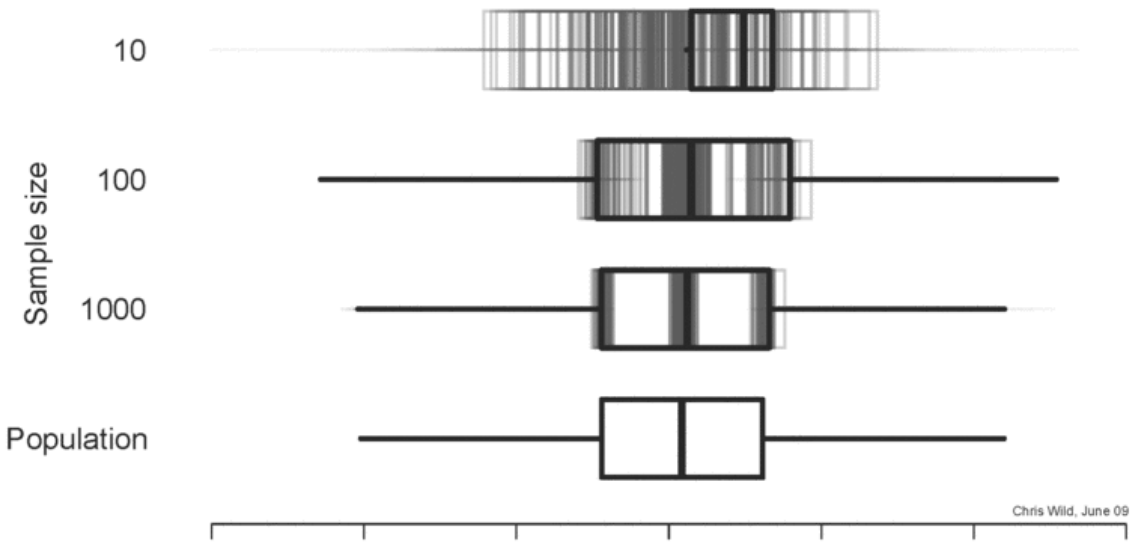
Variation in sample medians
Remember this diagram?

What is it telling you?

Every sample you take will have a different sample median. This is despite the population median remaining the same for all of your sample.

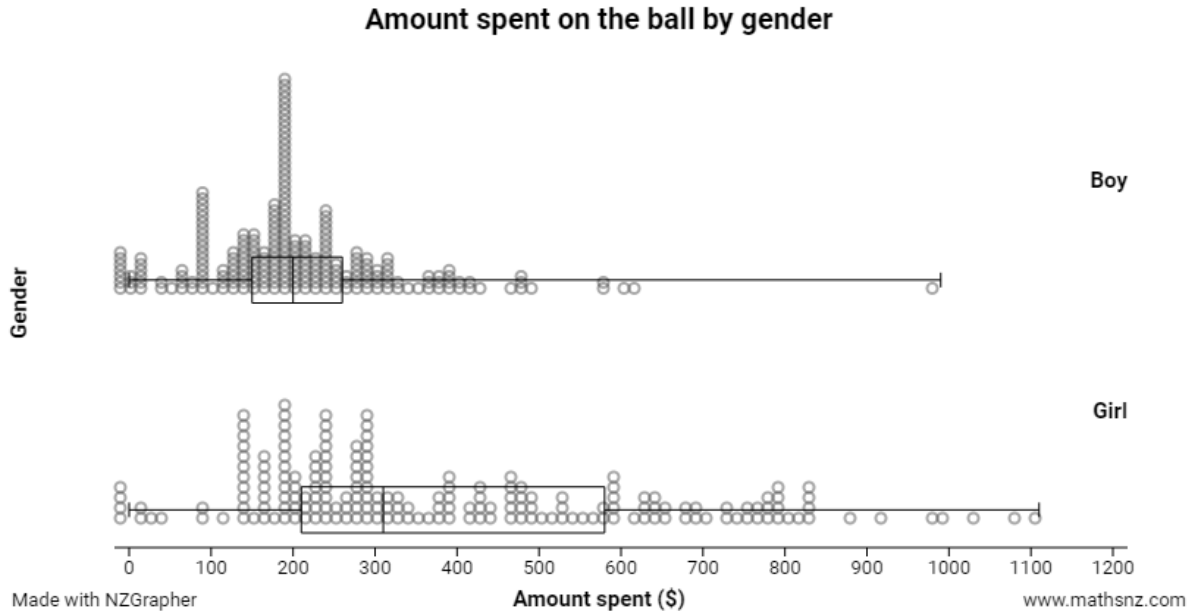
Pause and ponder that for a moment. Each sample gives a different sample median, but there has been no change in the population median. All of the "change" that you see, is due to sampling variation. This is the "noise" that obscures your view of the population.

Samples of larger sizes will have less sampling variation. For example, see the image below.



The interval in which we are "pretty sure" the population median is in shrinks as the sample size increases.

Conversely, if the population has a large spread, the median will have greater variation between samples so the interval will widen. For example, consider the following population:

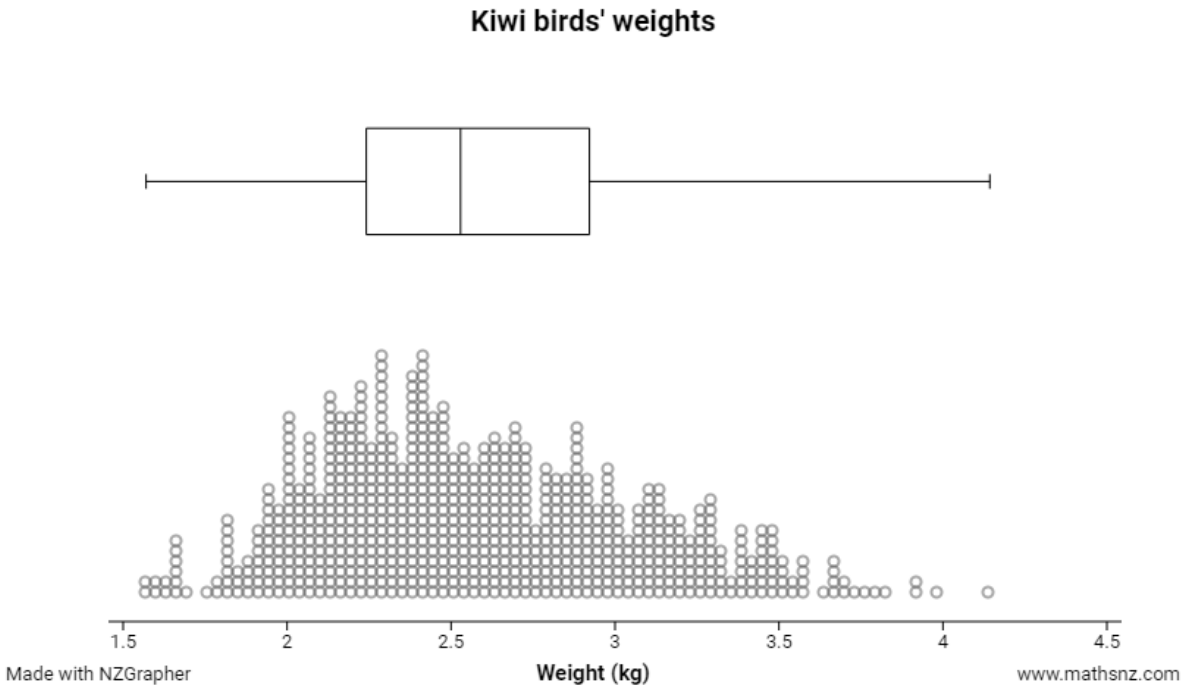


How does the spread compare between the two genders?

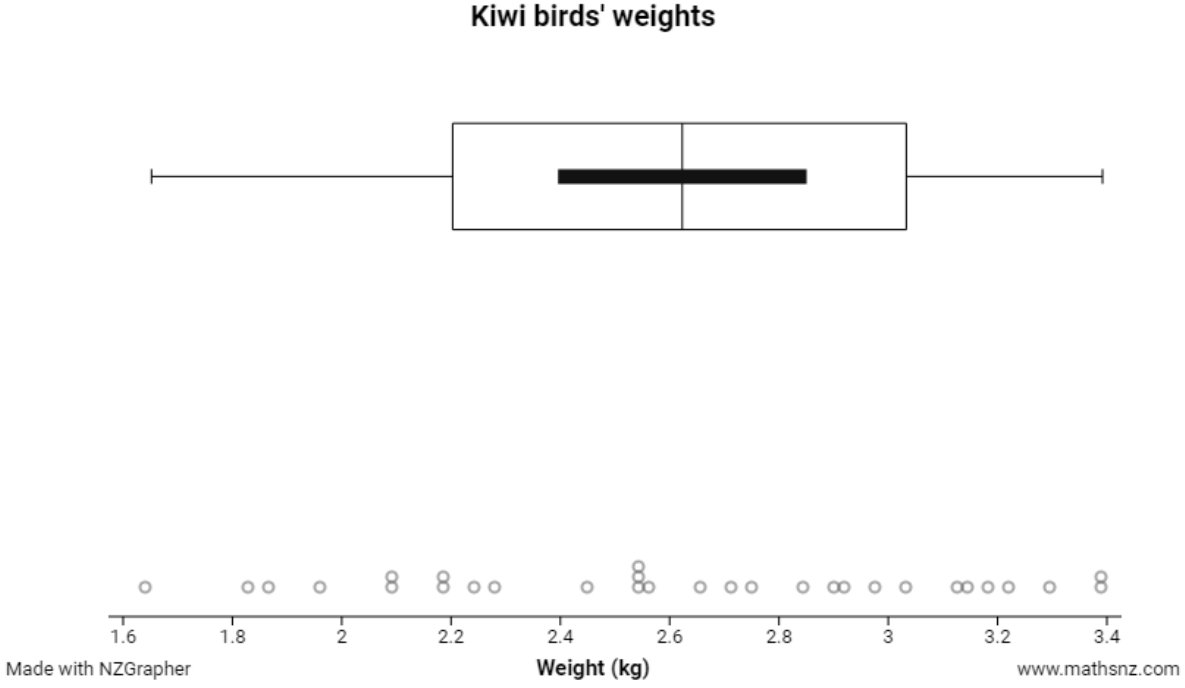
Watch the demonstration of the sample median variability. What do you estimate the intervals to be in which you are "pretty sure" the sample medians are for each gender?

Activity Five: Testing the 95% informal confidence intervals

Take a known population, say, all 700 kiwi in this dataset:



You can take samples from this population of, say, size 30 and create a 95% informal confidence interval like this:



How does the sample median compare to the population median?

_____

Does the 95% confidence interval contain the population median? _____

Watch the demonstration to see how often the 95% informal confidence interval contains the population median.