Level 2 – AS91264
Standard 2.9
4 credits – Internal

Use statistical methods to make an inference

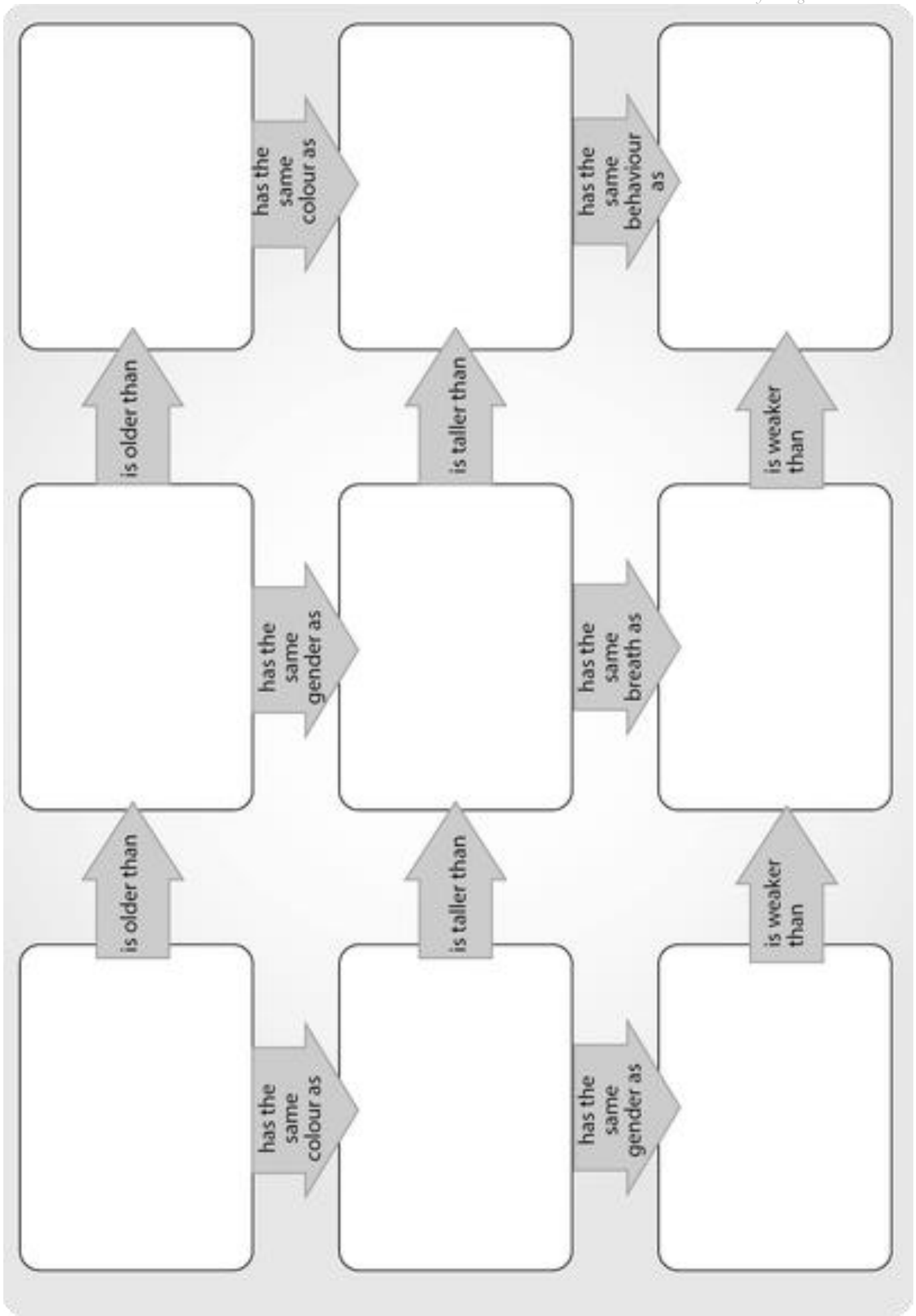
Activity: Dragonistics data cards

Introduction to the data cards
There is a population of dragons inhabiting an island. Each dragon has a number of attributes shown below:

Use nine data cards to fill in the table on the following page:

Then, 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.

Take some time to observe your dragon's characteristics. Write some hypotheses about the whole dragon population based on your sample, i.e. I think that there are more green dragons than red dragons or I think that fire breathing dragons are mainly female.


has the same colour as
has the same behaviour as
is older than
is taller than
is weaker than
has the same gender as
has the same breath as
is older than
is taller than
is weaker than
has the same colour as
has the same gender as

Lesson One: Investigative question (Problem)

An example from Dragonistics
Say you were trying to answer the question: "are green or red dragons stronger?" What issues are there with the wording of the question?

In 2.9 inference, you will compare some variable between two groups and investigate whether there is a difference, for example, are green dragons stronger than red dragons? But what exactly is meant by 'stronger'?

For this standard, you will investigate whether the median value for one group is greater than the median value for the other group in the population [by looking at only a sample].

For example, a suitable question would be: "I wonder if the median strength of green dragons is greater than the median strength of red dragons for the dragons on the island."

To be clear, a suitable question must contain:

- The numerical variable that is being investigated, e.g. strength.
- The groups being compared, e.g. green and red dragons.
- The population parameter you will be inferring about, i.e. the median.
- The population you are inferring about and from which you are sampling, e.g. all dragons on the island.
- The direction of the comparison, e.g. "Is the median strength for green dragons *greater* than…"

Write a suitable investigative question for the Dragonistics data set as the start of your practice internal.

A reminder about inferring
Remember that your investigative question must be about the entire population, e.g. all green and red dragons on the island. However, you only have access to a small sample of dragons from the island. Never forget that you don't have the full picture.

The above criteria are sufficient for an investigation [achieved] but to investigate with justification and/or insight [merit and excellence] you should include the following:

- Contextual background (e.g. "Adolescence is a time of great physical growth but different genders reach puberty at different times")
- A justification 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."*

If you wish, you could add context, justification, and your hypothesis to your investigative question for your practice internal.

Lesson Two: Sampling (Plan)

In your investigation, you are going to make an inference about the population medians for both groups, e.g. the median strength of all green dragons and red dragons on the island.

Unfortunately, you won't have data on all the dragons on the island, only some of them. This is called the sample.

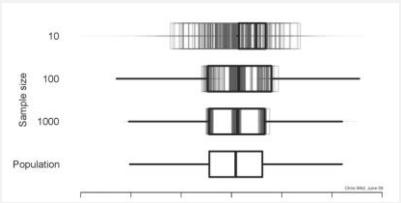
It's important that you have a good sample to make the best inference about the population. There are two considerations when taking a suitable sample.

Representative
A sample needs to reflect the population. A counter example, where the sample is unrepresentative, is if you chose mainly fierce dragons. In this case, we'd say that the sample is biased; it doesn't represent the population. It's reasonable to think that fierce dragons may have different strengths than other dragons so it is unsuitable to have a sample of mainly fierce dragons.

In order to have a representative sample (or in other words to eliminate bias), you should take a random sample from the population. In a random sample, each element of the population (dragon, person, etc.) has an equally likely chance of being selected for the sample.

Why is random sampling the ideal sampling method?

Size
All samples contain **sampling variation.** This is the variation that comes from sampling and causes samples to be different from one another and from the population.



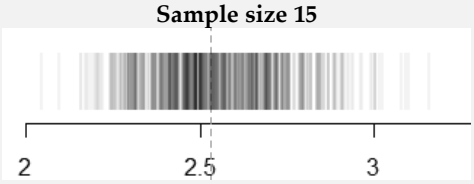
Remember when you each took samples of the Dragonistics cards in the first activity? Each of your samples had different median strengths and the median strength for all dragons was unknown, though you could guess what it would roughly be.

The larger the sample is, the less of an effect sampling variation has, though it will always have at least some effect.

It's important that the sample for **both groups** is big enough, i.e. enough green dragons and enough red dragons. It's a common misconception that the groups need to be the same size to be representative. They don't. They just need to both be big enough. For example, having 30 green dragons and 45 red dragons is large enough for both groups.

So what is 'large enough'? Well, that's up to you but here are some guidelines. If 100 different samples of size 15 were taken and all of the sample medians were recorded, the plot would look like this:

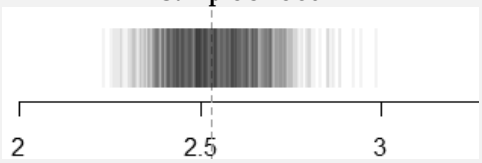
Sample size 15



This is a significant amount of sampling variation and won't allow a good inference to be made. Here, the population median could be anywhere between 2.2(ish) and 3(ish).

If samples of size 30 were taken and all of the sample medians were recorded, the plot would look like this:

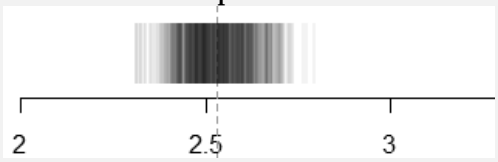
Sample size 30



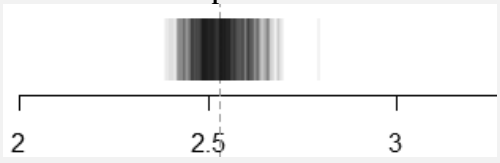
This is an allowable amount of sampling variation though not ideal.

Samples of size 60 and 120 have even less sampling variation:

Sample size 60



Sample size 120



Notice that there is not much difference between samples of size 60 and 120. After a point, there is less benefit to taking a larger sample. Given all of that, you choose what a suitable sample size is to minimise sampling variation without doing too much work.

Activity: Using NZGrapher (Data)

Making a box and whisker plot

Step One: Upload the data
Click "Choose file" in the top left corner and select the .csv file containing your data.

Step Two: Select graph type
Change the graph type to "Dot plot(and box and whisker)" in the bottom left corner.

Step Three: Select variables
Set variable one as your numerical variable, e.g. strength and variable two as your categorical variable (the groups you're comparing), e.g. dragon colour.

Step Four: Title and labels
Create an appropriate title for your graph. A good default is: "(Numerical variable) by (categorical variable)", e.g. "Strength by dragon colour" or "Marathon time by gender". Have suitable axis labels including units.

Step Five: Annotate graph
In the bottom centre section, tick the boxes to show:

- summaries
- box plots [or high box plots or box (no outlier)]
- informal confidence interval
- confidence interval limits

You may also want to show the mean dot and gridlines as well.

Step Six: Remove outliers (if applicable)
You may wish to remove some points as you consider them to be outliers. To do so, identify which row of the spreadsheet contains that point. You can do this by turning on point labels in the bottom centre section and selecting "Row +/-" in the top left corner and deleting the specific row that matches that data point. Make sure to turn off point labels before copying your graph.

Alternatively you can sort the spreadsheet by that variable by clicking "sample and more" in the top left corner. Sort by your numerical variable. Now you can remove the outlier by deleting the last (or first) row.

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)

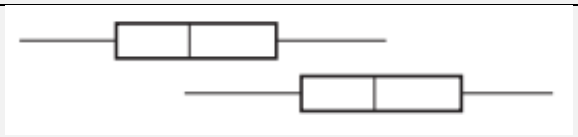
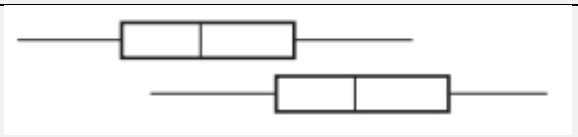
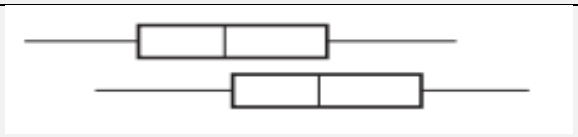
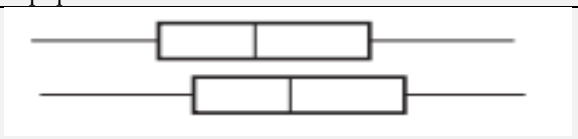
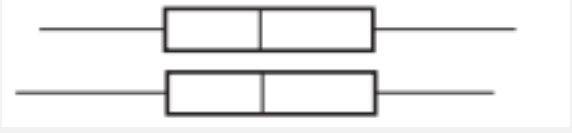
Often, these comments will flow back and forth between each other. They are not necessarily discrete and unrelated concepts.

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.

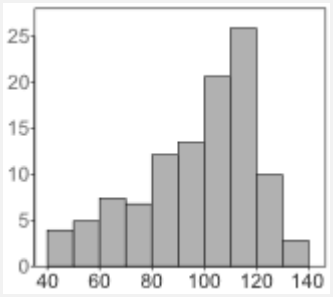
Shift and overlap
Shift describes the extent to which the data is "shifted" from one group to the other. The overlap describes the extent to which the data overlaps between the two groups. They are opposite measures; a large shift is the same as saying a small overlap and vice versa.

<table>
<tr><td> 
No overlap between the middle 50% for each group means a significant difference between the groups in the sample and is strong evidence for a difference in the population medians.</td><td> 
Both sample medians being outside the middle 50% for the other group also means a significant shift between the two groups in the sample and is strong evidence for a difference in the population medians.</td></tr>
<tr><td> 
One of the sample medians being outside of the middle 50% for the other group means there is a moderate shift between the two groups in the sample and is evidence for a difference in the population medians.</td><td> 
Both sample medians are inside the middle 50% for the other group means there is a small shift between the two groups in the sample. This could be evidence for a difference in the population medians depending on the sample sizes.</td></tr>
<tr><td> 
No difference in the sample medians means there is no shift between the two groups in the sample. It is impossible to say whether there is a difference in the population medians regardless of sample size.</td><td></td></tr>
</table>

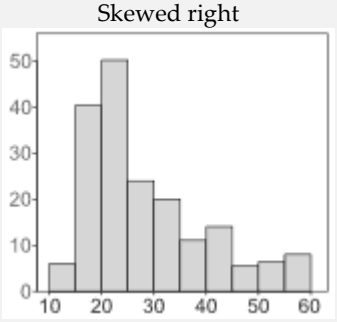
Shape

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

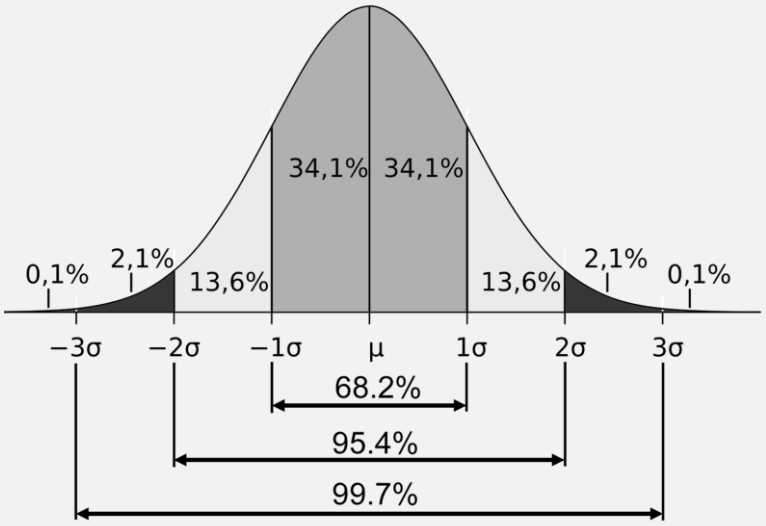


Spread

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 normally 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.

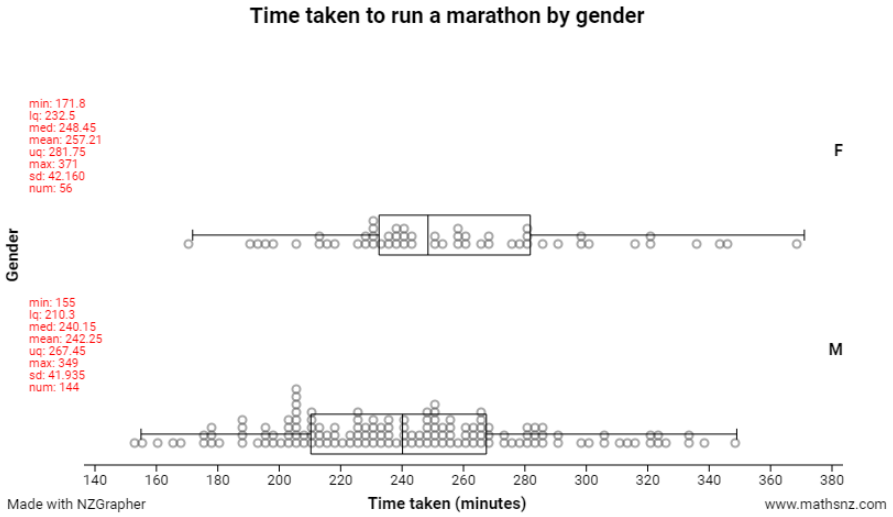
Signal vs. Noise

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?

The best estimate of the population median is the sample median. Though all estimates are wrong.

Explain sampling variability.

Making an 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.

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.

Grade requirements

	Achieved Use statistical methods to make an inference.	**Merit** Use statistical methods to make an inference, with justification.	**Excellence** Use statistical methods to make an inference, with statistical insight.
Problem	Pose a suitable investigative question. *Note: The investigative question that is posed must involve a comparison. This needs to include the variable, population groups being compared, population parameter (median) the inference will be about and the direction of the comparison.*	Justify the relevance of the investigation.	Show insight in the justification of the investigation.
Plan	Select random samples for each group. The sample method and sample size are stated.	As for achieved **plus the sampling method and/or sample size is justified.**	As for merit but **the sampling method and sample size are justified with insight.**
Data	Dot plot and box and whisker plot are produced with appropriate title and axis labels. Summary statistics and 95% informal confidence intervals are displayed for each group.		
Analysis	Compare at least two features of the distributions in context (i.e. referring to the variable name or giving units). This could involve comparing the center and middle 50%, shift and overlap, shape, spread, and unusual or interesting features.	Compare at least two features of the distributions in context **and justify the comparisons using evidence such as the visual features of the graphs or summary statistics.** **Link the discussion to the investigative question and the population (the "so what?" factor).**	Compare at least **three** features of the distributions in context with evidence and insight. *Insight could include contextual insight or statistical insight.*
Conclusion	The investigative question is answered by making an inference from the 95% ICIs. *An understanding of sampling variability and variability of estimates must be evident (this could be in the analysis).*	As for achieved, plus the 95% ICI is interpreted in context. The effect of at least one aspect, for example, sample size has been discussed (this could be in the analysis). *An understanding of the difference between the sample calculations and population estimates has been demonstrated.*	As for merit, plus key evidence from the analysis is summarised. An understanding of the context is shown by considering possible explanations for the findings.

2.9 Inference log

My goal for 2.9 Inference is: _____

Date	What did you do?	Did you work well?

Practice internal progress	Completed	To what level? (A/M/E)
Problem		
Plan		
Data		
Analysis		
Conclusion		