tūturu

Wellbeing Teaching and Learning Guide

NCEA Level 1

Mathematics and Statistics

1.10 Multivariate Inference

Internal Assessment Resource

Achievement Standard 91035

Investigate a given multivariate dataset using the statistical enquiry cycle

4 credits

Contents

Achievement Standard 91035

Subject Reference		Mathematics and Statistics 1.10			
Title		Investigate a given multivariate data set using the statistical enquiry cycle			
Level	1	**Credits**	4	**Assessment**	Internal
Subfield		Statistics and Probability			
Domain		Statistics			
Status	Registered	**Status date**	9 Dec 2010		
Planned review date	31 Dec 2020	**Date version published**	20 Nov 2014		

This achievement standard involves investigating a given multivariate data set using the statistical enquiry cycle.

Achievement Criteria

Achievement	Achievement with Merit	Achievement with Excellence
Investigate a given multivariate data set using the statistical enquiry cycle.	Investigate a given multivariate data set using the statistical enquiry cycle, with justification.	Investigate a given multivariate data set using the statistical enquiry cycle, with statistical insight.

1. This achievement standard is derived from Level 6 of *The New Zealand Curriculum* (2007), and is related to the material in the *Teaching and Learning Guide for Mathematics and Statistics*, Ministry of Education, 2010 at http://seniorsecondary.tki.org.nz. The achievement standard is aligned to the following achievement objectives taken from the Statistical Investigation thread of the Mathematics and Statistics learning area:
    - carry out investigations of phenomena, using the statistical enquiry cycle:
        - determining appropriate variables
        - cleaning data
        - using multiple displays, and re-categorising data to find patterns, variations, in multivariate data sets

- comparing sample distributions visually, using measures of centre, spread, and proportion
- presenting a report of findings

- plan and conduct investigations using the statistical enquiry cycle:
  - justifying the variables used
  - identifying and communicating features in context (differences within and between distributions), using multiple displays
  - making informal inferences about populations from sample data
  - justifying findings, using displays and measures.

This standard is also derived from *Te Marautanga o Aotearoa*. For details of the *Marautanga* achievement objectives to which this standard relates, see the Māori version of the standard.

2. *Using the statistical enquiry cycle* involves using each component of the statistical enquiry cycle to make comparisons.

   Using the statistical enquiry cycle with justification involves linking aspects of the statistical enquiry cycle to the context and the population and making supporting statements which refer to evidence such as summary statistics, data values, trends or features of visual displays.

   Using the statistical enquiry cycle with statistical insight involves integrating statistical and contextual knowledge throughout the statistical enquiry cycle, and may involve reflecting on the process or considering other explanations for the findings.

3. Students need to be familiar with the statistical enquiry cycle to investigate a given multivariate data set, which involves:
   - investigating data that has been collected from a survey situation
   - posing an appropriate comparison question using a given multivariate data set
   - selecting and using appropriate display(s)
   - giving summary statistics such as the five summary values (minimum, maximum, median, quartiles)
   - discussing features of distributions comparatively, such as shape, middle 50%, shift, overlap, spread, unusual or interesting features
   - communicating findings, such as informal inference and supporting evidence, in a conclusion.

4. Conditions of Assessment related to this achievement standard can be found at http://ncea.tki.org.nz/Resources-for-Internally-Assessed-Achievement-Standards.

New Zealand Curriculum Achievement Objectives

Achievement Objective S6-1

In a range of meaningful contexts, students will be engaged in thinking mathematically and statistically. They will solve problems and model situations that require them to:

- plan and conduct investigations using the statistical enquiry cycle:
  - justifying the variables and measures used
  - managing sources of variation, including through the use of random sampling
  - identifying and communicating features in context (trends, relationships between
  - variables, and differences within and between distributions), using multiple displays
  - making informal inferences about populations from sample data
  - justifying findings, using displays and measures.

Introduction

This unit of work gives students the opportunity to work at Level 6 of the NZ Curriculum, focussing mainly on Achievement Objective S6-1, which is primarily aimed at using the statistical enquiry cycle with a given multivariate dataset.

The unit focusses on our generic heading of 'Wellbeing', which is a term used to describe the level of wellness or condition of an individual. While the oxford dictionary defines wellbeing as 'the state of being comfortable, healthy, or happy', there are many variables that contribute to a positive or negative level of wellbeing;

- home environment and support from home
- school environment and support whilst at school
- community links and communal activities such as church, sports etc
- strong links to friends and family
- substance abuse, if any

Measuring wellbeing is challenging as there are a number of different variables to take into account. It can be done in a number of ways and there is no 'one size fits all' approach. The largest study of its kind in New Zealand, has been recently completed in 2012 by the Adolescent Health Research Group, in an attempt to look into the wellbeing of young people in New Zealand aged between 13 and 17. The data has been made available for discussion and analysis and you will be exploring this data in this unit of work.

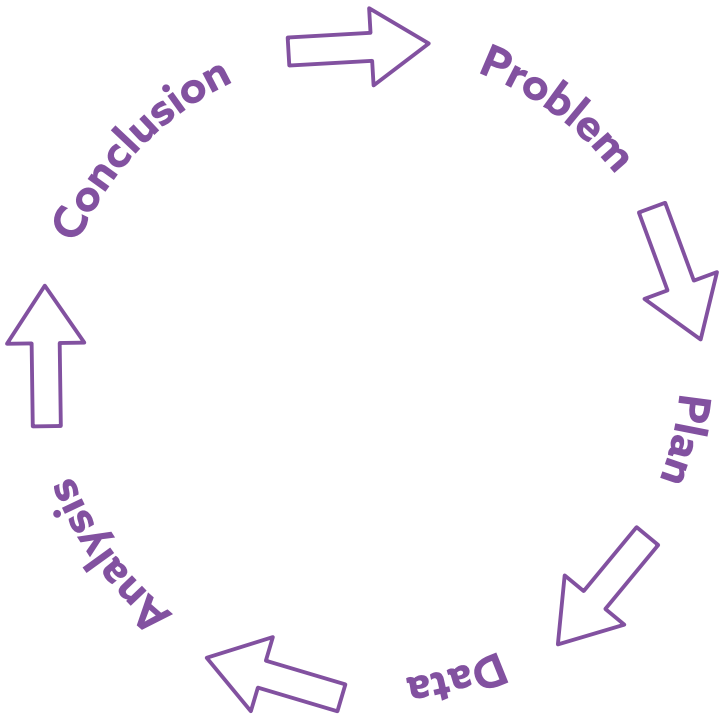
This unit focuses on a few of the variables from this study in an attempt to see a snapshot of the lives of young people aged 13 to 17 in New Zealand.

Questions

1. "This unit focuses on a few of these variables in an attempt to see a snapshot of the lines of young people ages 13 to 17 in New Zealand."
   a. **What do you think the word 'snapshot' means in this sentence?**
   b. **What other variables can you come up with, that would contribute to a positive or negative 'wellbeing' level?**

The Statistical Enquiry Cycle

The statistical enquiry cycle is a cycle that is used to carry out a statistical investigation. The cycle consists of five stages: Problem, Plan, Data, Analysis, Conclusion. The cycle is sometimes abbreviated to the PPDAC cycle.



1. Problem

The problem section is about formulating a statistical question, what data to collect, who to collect it from and why it is important. What is it that you want to measure?

2. Plan

The plan section is about how the data will be gathered. Look at how you can solve the problem. How are you going to measure it? Collect the data.

3. Data

The data section is about how the data is managed and organised. Clean and sort the data.

4. Analysis

The analysis section is about exploring and analysing the data, using a variety of data displays and numerical summaries, and reasoning with the data. Make tables and graphs. Look for patterns. Explore what is going on.

5. Conclusion

The conclusion section is about answering the question in the problem section and giving reasons based on the analysis section. Answer the question. Explain what this means. This may lead to other questions.

1. Problem

In this standard you are making a comparison between two different groups in the population. The problem is a comparison question (comparing one group with the other) and needs to be worded to include:

- a clearly identified population
- what groups are being compared
- what variable is being investigated
- an indication of central tendency, i.e. using 'tends to'
- the direction of the comparison, i.e. greater than or less than

Example

I wonder if the time teenagers spend on mobile phones ***tends to be*** *longer than the time adults spend on mobile phone from an online study in Australia in 2015.*

The **population** identified in this investigation is those who participated in the online study in Australia in 2015

The **groups** being compared are teenagers and adults

The **variable** being investigated is the time spent on the mobile phone

Tendency has been shown by using the words *tends to*. Other words could be used such as 'on average' – as long as it is clear from the question we are not saying incorrectly that *all* teenagers spend more time than *all* of the adults.

The **direction** of the comparison is that teenagers tend to be longer.

Questions

1. I wonder if the wellness score for female students tends to be greater than the wellness score for male students who took part in the New Zealand Youth Health and Well-being survey in 2012.

   a. **What is the population identified in this investigation?**

   b. **What are the groups being compared?**

   c. **What is the variable being investigated?**

   d. **What is the direction of the comparison?**

Answers:

a. *those who participated the New Zealand Youth Health and Well-being survey in 2012*

b. *females and males*

c. *the wellness score*

d. *whether the female score tends to be greater than the male score.*

2. Plan

In the plan you need to include the following in your report:

- discuss what you intend to do in the investigation.
- the variables being investigated including units
- discuss where the data has come from

Samples From the Population

A **population** consists of all elements from a set of data where as a sample contains only a part of a population. While the most accurate conclusions can be drawn from having access to all the data from the population, this is often not possible. Therefore a sample is taken and from that sample, we can get an idea of what is likely to be happening in the population.

Samples are traditionally gathered through conducting a survey which can be through observations, interviews or questionnaires. In the very online world of today data for samples can be gathered through online tracking and social media monitoring without people even being aware of it happening.

A large **sample size** ensures that the data is **accurate** enough to represent the population.

The New Zealand Youth Health and Well-being Survey 2012

In 2012, a two-stage random sample was completed to collect a survey population of 2,996 students.

Stage One

Firstly 125 schools were randomly selected throughout the country to partake in the survey.

Stage Two

Of schools that opted into the survey, 20% of the students were randomly selected to complete the survey. If the school had less than 150 students, 30 students were selected to ensure a viable group has been selected and also to ensure privacy was at its maximum.

These students were invited to participate in the New Zealand Youth Health and Well-being Survey 2012, one in which parental permission had to be obtained, and

of which only 2,996 students responded. The population for this investigation is the Participants of the New Zealand Youth Health and Well-being Survey 2012. Students completed a questionnaire on a hand-held tablet, allowing questions to be presented in an audio-visual form.

In this standard, you will be provided with a random sample of 200 students from within this population, to complete your statistical investigation.

Questions

1. "Firstly 125 schools were randomly selected throughout the country..."
   a. **Would this sampling technique exclude any young person aged between 13 and 17 throughout New Zealand?**
   b. **How would schools take a sample of 20% of their roll?**

3. Data

Data may need to be cleaned to remove any inappropriate or missing values. Before using a data set:

- check that it contains the data we expect
- clean the dataset by removing blanks or unexpected values

The New Zealand Youth Health and Well-being Survey 2012 Dataset

Variable	Description
Gender	Male or Female
Age (years)	The age of the student rounded down to the closest whole year
Wellbeing Score	This variable is the World Health Organisation-5 Wellbeing Index (WHO-5). Scores can range from between 0 to 25. The larger the WHO-5 score, the more likely they felt cheerful, calm, active, rested, and had things in their life that interested them over the previous two weeks.
Feelings	This variable is called the RADSSF score. Scores can range from between 10 and 40. The larger the RADSSF score, the more likely the student experienced low moods.
Challenges	This variable is the Total Difficulties Score from the Strengths and Difficulties Questionnaire. Scores can range from between 0 to 40. The larger the score, the more likely they had experienced social and emotional challenges over the past six months.
School Connection	This variable is the school connection score. Scores can range between 0 to 4. The larger the score, the more likely they had felt connected to their school.
Family Connection	This variable is the family connection score. Scores can range between 0 to 4. The larger the score, the more likely they had felt connected to their family.

Family Meals	This variable is the response to the question: "*During the past 7 days, how many times did all, or most, of your family living in your house eat a meal together?*" Responses were given the following values: Never = 1 1-2 times = 2 3-4 times = 3 5-6 times = 4 7 or more times = 5
Binge Drinking	This variable indicates that the student reported binge drinking once or more in the last month. Binge drinking was defined as 5 or more alcoholic drinks within a 4 hour period. Not at all = 0 Once or more = 1

Questions

1. Researchers have spent the last few years analysing the data ready for release to the public.
   a. **The study was conducted in 2012. Would there have been any / many changes since then?**
   b. **Which of these variables do you believe would be changing the greatest for young people living now?**

2. One of the groups that could be used in the investigation is gender;
   a. **How critical is gender in this investigation?**
   b. **How much different are teenage males to teenage females, with these variables in mind?**
   c. **Which of the variables would be greatest (or least) affected by gender?**

4. Analysis

Once the groups and variables have been identified and the data collected from the sample, the data needs to be analysed.

You will need to;

Use appropriate displays:

- Dot Plot and/or
- Box and Whisker Graph

Calculate appropriate sample statistics:

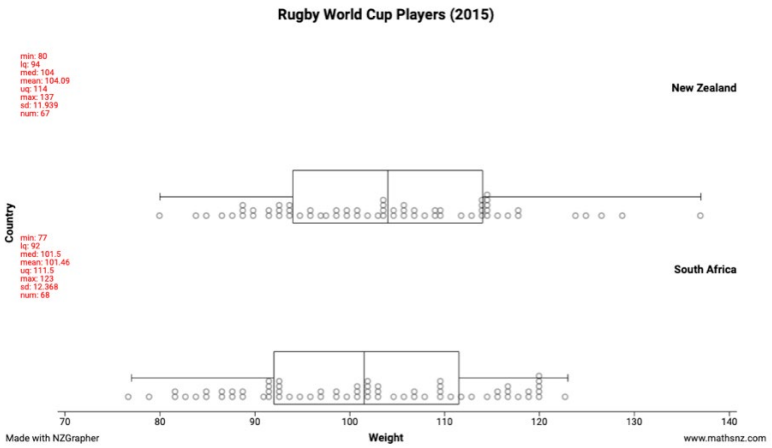
- Central Tendency (mean, median)
- Spread (max, min, range, lower quartile, upper quartile, IQR)

Discuss and compare the sample distributions. It is important that you compare two groups. It is not enough to simply describe the features of the two groups separately:

- Centre
- Shape
- Spread
- Unusual Features

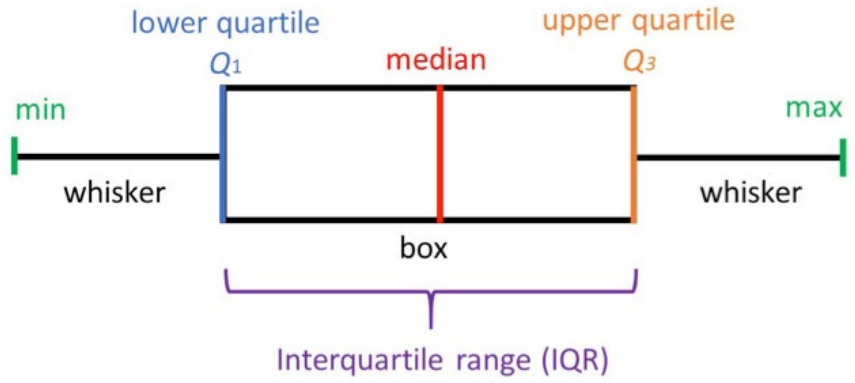
Display

The most useful graphical display for this standard is the dot plot with box and whisker graph. There are a number of suitable software packages that can do this - popular are the online resources iNZight and NZGrapher. The dot plot with box and whisker graph below has been created with NZGrapher. It shows the summary statistics on the left.



Box and Whisker Plot

Box and Whisker plots (also known as just a box plot) are used to visually show the distribution of data through displaying the minimum, lower quartile (LQ), median, upper quartile (UQ) and maximum values of the dataset.



Parts of the Box and Whisker Plot

1. **Minimum**
   The lowest value in the dataset.

2. **Lower Quartile (LQ)**
   Twenty-five percent of values fall below the lower quartile value. This marks the start of the box.

3. **Median**
   The median marks the mid-point of the data and is shown by the line that divides the box into two parts. Note that the median line may not be exactly in the middle of the box – it could be closer to the LQ or closer to the UQ depending on the skewness of the data.

4. **Upper Quartile (UQ)**
   Twenty-five percent of the values are above the upper quartile value. This marks the end of the box.

5. **Maximum**
   The highest value in the dataset.

6. **The Interquartile Range (IQR)**
   Between the UQ and the LQ is the box showing the middle 50% of scores. Its size is determined by UQ minus the LQ

Discuss and Compare the Sample Distributions

Feature	Discussion
Centre	Compare the centre of each group by identifying the medians and saying which is greater. Justify by calculating the difference so you can show how much they differ by.
Shape	You will need to compare by describing the overall shape of the data for each group. Use the following keywords to describe: ▪ Normal distribution (symmetrical, bell shaped curve) ▪ Left skewed (tail on the left side) ▪ Right skewed (tail on the right side) ▪ Multi modal (more than one peak) ▪ Uniform (rectangular shaped, generally all one level) Justify the shape chosen by thinking about the following features: ▪ symmetry ▪ the tail that appears either end ▪ how many peaks are visible
Spread	Compare the spread of the data in each group by using any of the following: ▪ the range (distance from the minimum to the maximum data values) ▪ the interquartile range IQR (distance from the lower quartile to the upper quartile) ▪ whether the boxes overlap each other or not Justify by calculating the difference so you can show how much they differ by.
Unusual Features	If there is anything unusual in the data from either group, point this out. ▪ outliers (datapoints that lie out at the extremes) ▪ large clusters that stand alone ▪ significant gaps that may be apparent in the data

5. Conclusion

Clearly communicate your findings in a conclusion, linking your findings to the context and population. A conclusion looks back to the Problem and attempts to answer the investigative question and this involves making an inference about the population.

Your conclusion needs to include the following:

- Make a supported inference about the population – 'making a call'
- Discuss sampling variability, including variability of estimates.

You may also wish to include the following:

- link back to your prediction
- how could the investigation be extended?
- what could this inference lead on to next?

Inference

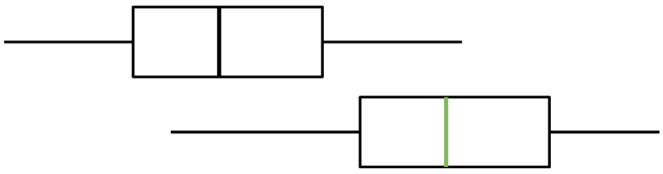
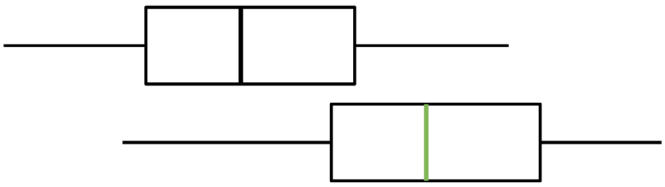
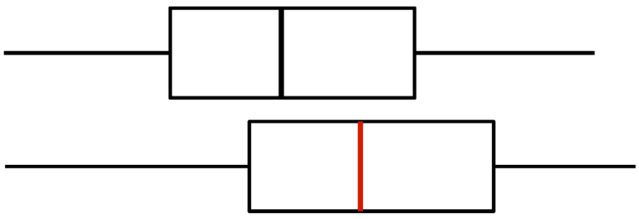
When you answer the investigative question you are making an **inference**. You are using the information you have gathered from the data in the sample to make an inference about the population. Your answer therefore is relevant to the specific population you are interested in so it is important that you are clear about what that population is. For example data gathered from a wellness survey in America in 1980 cannot be used to answer questions that we might have about wellness in New Zealand today. Not only is the time period significantly different but the demographics, lifestyle choices, cultures and opinions etc are all very different to those from the New Zealand population.

Because your conclusion is based on one sample it is possible that another sample could produce different results and this is known as **sampling variability**. Larger samples will contain less sampling variability and consequently lead to a more accurate inference being made. For example a sample size of 10 people taken from the same population of 1,000 will give you a very different result than a sample size of 200.

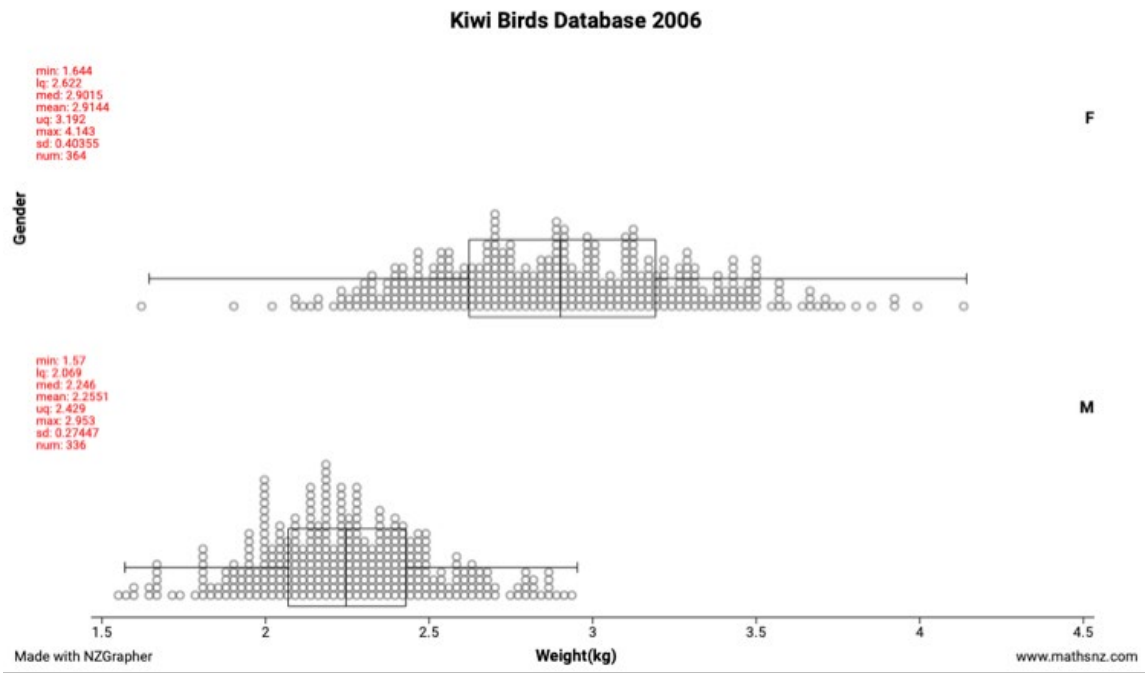
Making the Call

Look at the box and whisker plot and use this to decide if there is enough evidence to say if there is a large enough difference between the two medians.

- If the boxes don't overlap, or if the boxes overlap but the median is not yet inside the box of the other then the medians are far enough apart so that you can make the call.
- If the boxes overlap and the medians are inside the box of the other you can't make the call

Making the call	Position of box and whisker graphs
✓ Can make the call	 There is no overlap of the boxes and therefore the medians are clearly outside the box of the other.
✓ Can make the call	 There is overlap of the boxes but the medians still lie outside the box of the other. (Note: *either* one *or* both medians outside is good)
✗ Can't make the call	 There is overlap of the boxes and the medians lie inside the box of the other.

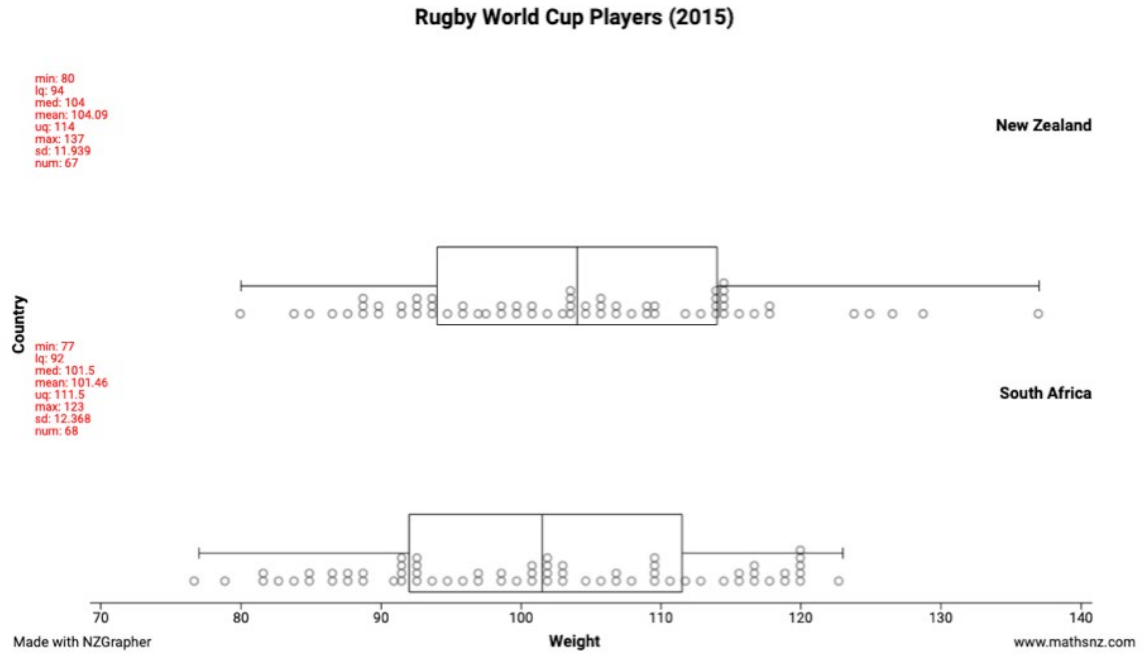
Example



*"For the population of kiwi birds from the kiwi bird dataset 2016, we **can** make the call that female kiwi birds tend to be heavier than male kiwi birds because the boxes do not overlap ."*

Questions

1. From the following sample of Rugby World Cup players displayed below, can we make a call that New Zealand rugby players tend to be heavier than South African rugby players?



**Answer: We can't make the call that New Zealand rugby players tend to be heavier than South African rugby players from the Rugby World Cup 2015 because the median of the South African rugby player weight lies inside the box of the New Zealand rugby players box, and the median New Zealand rugby player weight lies inside the box of the South African rugby players box.*

1. "Students have been provided with a random sample of 200 students from within this population, to complete the statistical investigation"
   a. **If the 200 students in your sample is a good representation of all 2,996 students that completed the survey, could your inference be a good representation of the survey population?**
   b. **Could it be a good representation for all young people aged 13 to 17 in New Zealand? Why or why not?**

Suggested Teaching Sequence

The teaching sequence for this unit of work has three phases:

<table>
<tr><th>Phase</th><th>Time</th></tr>
<tr><td>Phase 1</td><td>2-3 weeks</td></tr>
<tr><td colspan="2">Students refresh prior learning and develop the Level 7 understandings that will be assessed by Achievement Standard AS91264 (Use statistical methods to make an inference);

They should do this in a variety of contexts other than student wellbeing. The following is a suggested outline:

▪ Statistical Enquiry Cycle review (PPDAC):
▪ Posing a comparison question
▪ Planning
▪ Data: cleaning & collection
▪ Analysis: summary statistics and sample distributions:
- centre, shape, spread, overlap
- special or interesting features.
▪ Overlap
▪ Conclusion
- Inference and ‘making the call’
- sampling variability in relation to the inference</td></tr>
<tr><td>Phase 2</td><td>1 week</td></tr>
<tr><td colspan="2">Set a formative assessment task for learners to work through and use this as a vehicle to check their understanding of the key ideas and to provide specific feedback. The context for this task should be relevant and meaningful to the learners. Support should be provided for students during this practice phase, which provides an opportunity to identify and address any misconceptions or areas of weakness.</td></tr>
<tr><td>Phase 3</td><td>1-2 weeks</td></tr>
<tr><td colspan="2">Students are introduced to the context of student ‘wellbeing’. Students work with the data set to:
▪ clarify what the variables represent
▪ select the variables they want to work with
▪ pose an appropriate investigative comparison question.
Students complete and submit the assessment task.</td></tr>
</table>

Help Organisations

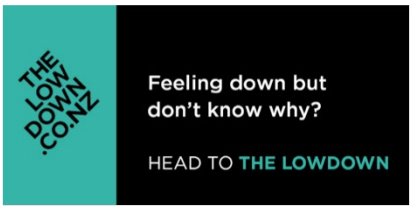
Help for the Tough Times

Help for the Tough Times provides a quick guide to four New Zealand websites that were especially designed to support teens with issues like anxiety, stress, identity, relationships, and depression. A pocket guide was co-developed for teens by teens (by two Year 12 high school classes). Support material is also available for school staff so that they are aware of the websites.

Order FREE hard copies from this website
www.hpa.org.nz/education/help-for-the-tough-times

The Lowdown

www.thelowdown.co.nz
Free phone 0800 11 17 57 or free text 5626



Youthline

www.youthline.co.nz
Free phone 0800 37 66 33 or free text 234

AlcoholDrug Helpline

www.alcoholdrughelpline.org.nz

AuntyDee

www.auntydee.co.nz

Sparx

www.sparx.org.nz

RainbowYOUTH

www.ry.org.nz

Mental health and addictions helpline

Free phone 1737 and free text 1737