Student Booklet

Te whai hua - kia ora!

sorted in Schools

Statistical Inference

LEVEL 2

CREDITS 4

AS91264 (version 3)

Mathematics and Statistics

Use statistical methods to make an inference
Te whai i ngā tikanga o te tūhuratanga tauanga hei whakaputa hīkaro

SORTED THEMES

KiwiSaver, Retirement, Managing my Money

Contents

Nau mai haere mai!

Welcome to the Statistical Inference module.

Please read through the Student Guide for an overview of the module and assessment before starting this Student Booklet.

1

Topic One:
Establishing a purpose and an investigative question

Learning outcomes for Topic One

- ✓ Understand what statistical inferences are
- ✓ Understand the purpose of an investigation
- ✓ Understand the key components of an investigative question; population, subgroups, variables, and measures.

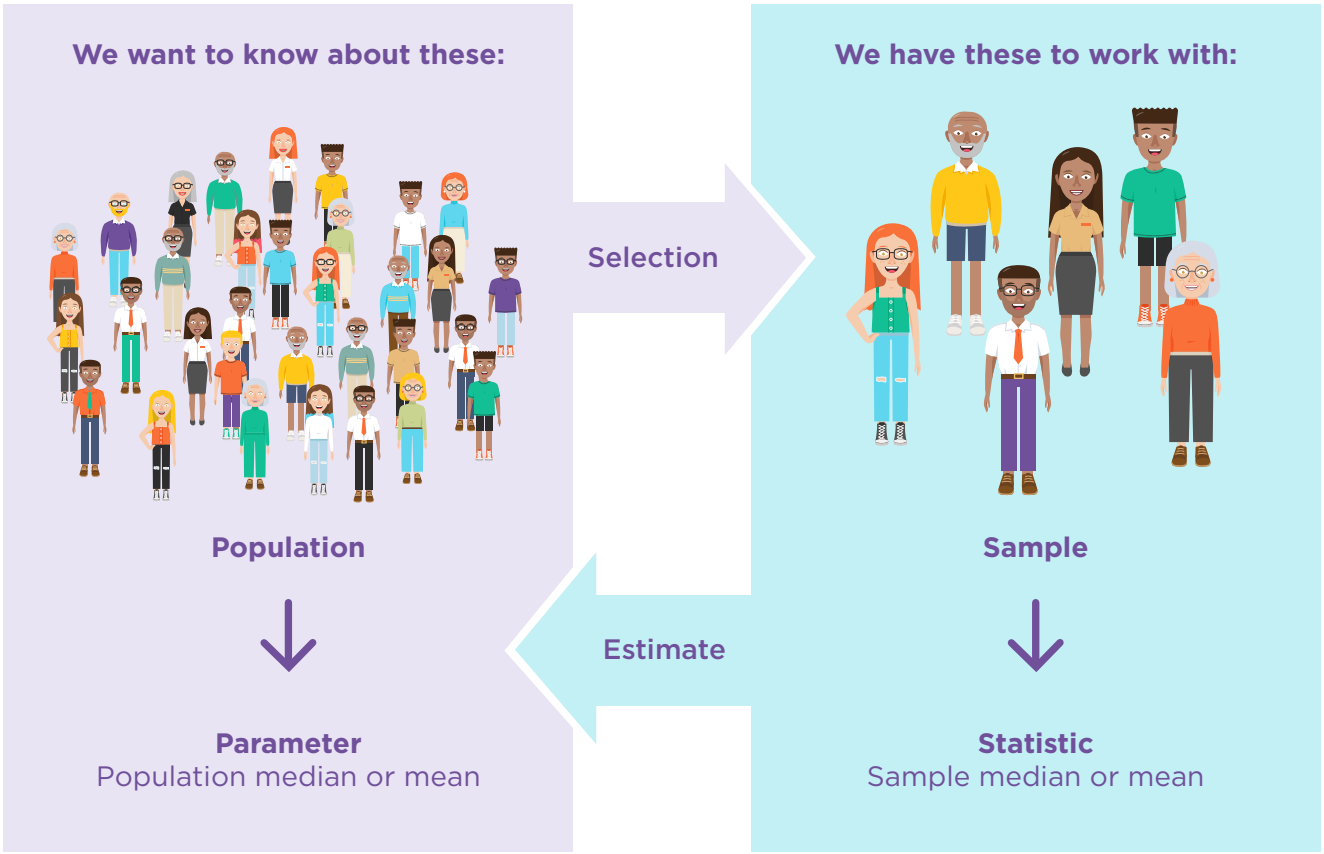
Success criteria

- I can describe the purpose of an investigation
- I can explain what a population is
- I can explain the difference between a categorical and a numerical variable
- I can explain what the median and mean are
- I can explain the difference between a parameter and a statistic
- I can write an investigative question.

The focus of this module is on learning how to use statistical methods to make an inference.

Making a statistical inference involves taking a sample from a population, analysing it, and then using the sample to make an informed guess about the population it came from. This diagram provides an overview of the process:

Statistical Inference



Why are inferences useful?

Imagine that you wanted to find out the average income of everyone aged 15 and older in Aotearoa New Zealand. In a statistician's ideal world, this would be done by contacting everyone in the population, asking each person a question in a language they can understand, and being certain that the answer they gave is accurate. But let's think about this in real terms. In 2020, the population of New Zealanders aged 15 and older was around 4,100,000 people. If 100 people each worked 8 hours a day, successfully calling a new person every two minutes, it would still take 170 days (nearly 6 months) to contact everyone.

In that timeframe, some people's incomes would have gone up or down, so even with that mammoth effort, you still couldn't be sure that you had the exact amount.

Although an actual average income exists, it's impossible to find out what it is. Fortunately there's another option, which is to take a large group from the population, find out what their average income is, and use this to estimate the average income of the population.

This process is called "making an inference", and throughout this module you'll be learning how to use statistical methods to make inferences yourself. (Statistical methods are mathematical formulas, models, and techniques that are used to analyse data.)

Two key concepts underpin this module.

1) **A sample is never an exact match for the population it comes from.** This means that the median or mean of the sample is unlikely to be exactly the same as the actual population median or mean. It can only ever be an estimate.

2) **No two samples taken from a population are exactly the same.** If two people each take a sample from a population and then use their results to make an inference, they will get different results.

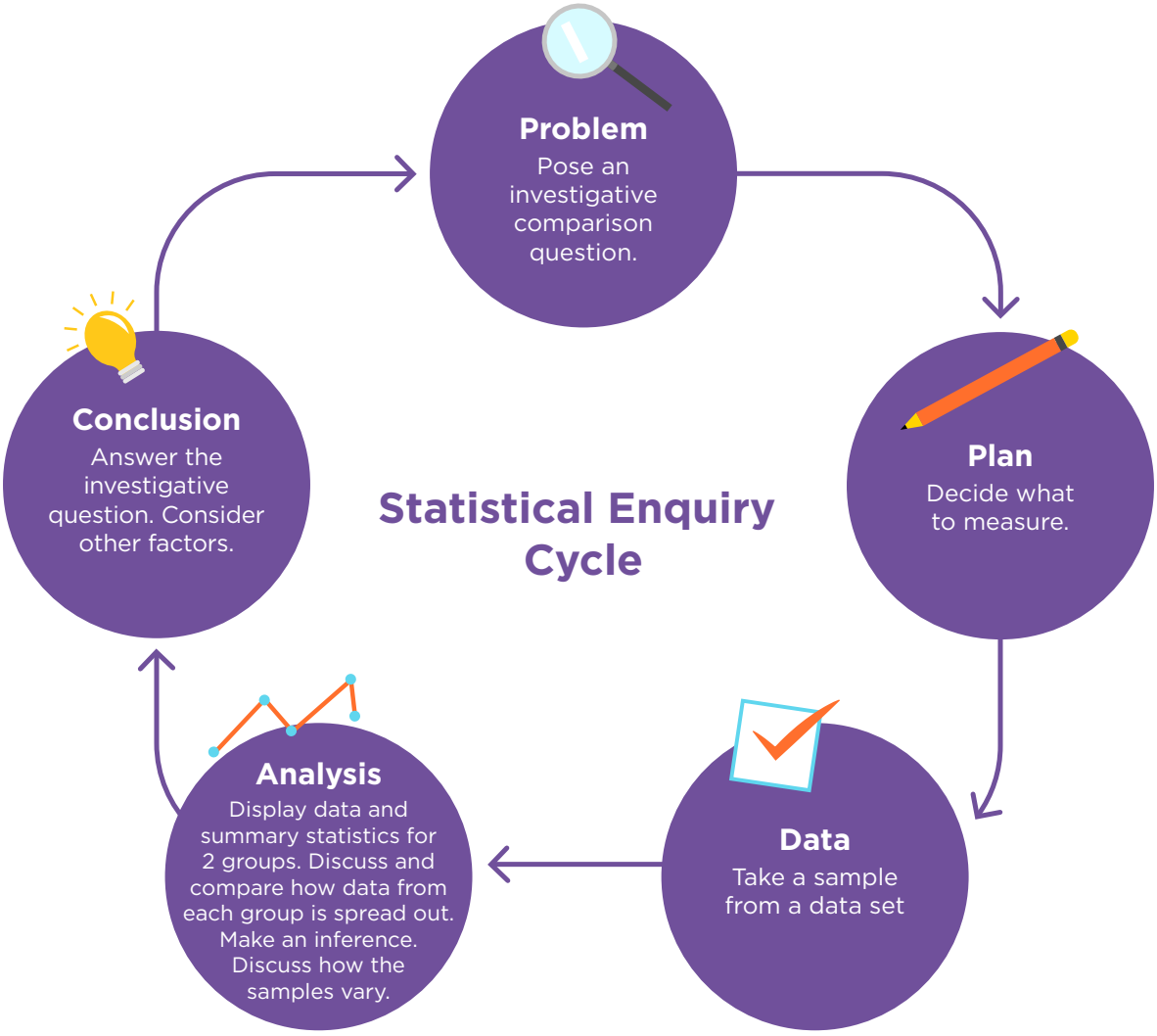
To reflect the inaccuracy and variability of samples, statisticians usually provide a set of values that they believe contains the population median or mean rather than just stating one value. You'll be using software to do this too.

Assessment tip:

Showing that you understand that samples always have some degree of inaccuracy and that no two samples are exactly the same is an essential part of your assessment.

The statistical enquiry cycle

The process of making an inference involves each component of the statistical enquiry cycle:



Identifying the purpose of an investigation

The starting point for any statistical investigation is identifying a problem and developing an investigative question.

The area explored in this module is:

Does the gender pay gap influence how much money people invest in KiwiSaver?

For background information on both the gender pay gap and KiwiSaver, see the Statistical Inference Student Guide.

Complete Topic 1 Activity 1
in the Student Practice Booklet

Assessment tip:

In your assessment, you need to explain why the purpose of your investigation is useful and/or interesting.

Why is it useful to investigate the KiwiSaver investment gap?

Throughout the world, females face challenges saving for retirement. In Aotearoa New Zealand, a **2018 research paper** found that twice as many females aged 65 and over are living in poverty compared to males aged 65 and over (14 percent compared with 6.6. percent). 75 percent of females stop contributing to KiwiSaver when they have children. The Commission for Financial Capability is exploring whether females who take time out of paid work should have a "care credit" paid into their KiwiSaver accounts as a way to help close the gender investment gap in Aotearoa New Zealand. Being able to provide evidence about females having lower KiwiSaver balances than males of the same age would strengthen the argument for a "care credit" to be introduced.

To track and address the gender investment gap, we need up-to-date information about how much money people have invested in KiwiSaver. There's no straightforward way to get this information from the entire population, which is why we need to make inferences about the amount that females and males have invested in KiwiSaver based on a sample.

To explore this problem, we need to create an investigative question that will guide the data we collect, and allow us to analyse and draw a conclusion from it.

Investigative questions have several key components:

- the population being investigated
- the subgroups of the population we want to compare
- the variable we are interested in
- the measure we are using for our investigation.

Each of these components is explained on the following page.

Financial status of New Zealanders aged 65 and over

Defining the population

In statistics, a **population** means an entire group of people, animals, plants, or things that we want to learn about or describe. When you define a population of interest, you need to be as specific as you can, including defining the timeframe and place.

In this module, the populations being investigated are "male and female New Zealanders aged between 55 and 64 in 2018".

Why this population?

People aged 55 to 64 are getting close to retirement age and will have fewer years to prepare financially for retirement. Looking at this group of people will reveal whether there is a gender investment gap that is likely to affect these two groups in retirement.

The date is based on when the data we will be using was collected. **You can access the CFFC Barometer data here**.

Choosing sub-groups from the population to compare

To align with the problem we are investigating, the two groups we will be comparing are:

- females in Aotearoa aged 55 to 64 in 2018
- males in Aotearoa aged 55 to 64 in 2018.

Note that ideally the problem would include people who are gender diverse, but in the data set we are using, only two people in the data identified themselves as gender diverse. This group is too small to include in the investigation.

Defining the variables of interest

A **variable** is the aspect of the population we want to describe.

There are two main types of variables:

- categorical variables; variables that describe qualities or characteristics, for example, the region that you live in, your gender, or your ethnicity
- numerical variables; variables that are based on counting or measuring something, for example, how much you spend each week.

Complete Topic 1 Activity 2
in the Student Practice Booklet

If we want to investigate whether males or females have more money invested in KiwiSaver, then the variable we will be using is **the amount of money a person has invested in KiwiSaver** (a numerical variable), which is presented in the **KiwiSaver balance** column of the CFFC Barometer data set.

Complete Topic 1 Activity 3
in the Student Practice Booklet

Assessment tip:

In your assessment, you'll be given a set of data to use as the basis of your investigation. The data set will necessarily limit your choice of population, sub-groups, and variables. It's worthwhile spending some time familiarising yourself with the data set before you decide what you are going to focus on.

Measures of interest

The next step is to decide which measure we will use to compare the two groups we have chosen. A measure is a value that describes a group. Two of the most commonly used measures are the median and the mean.

The **median** is worked out by putting all of the data values in order and finding the middle value.

The **mean** is calculated by adding a set of values together then dividing by how many values there are.

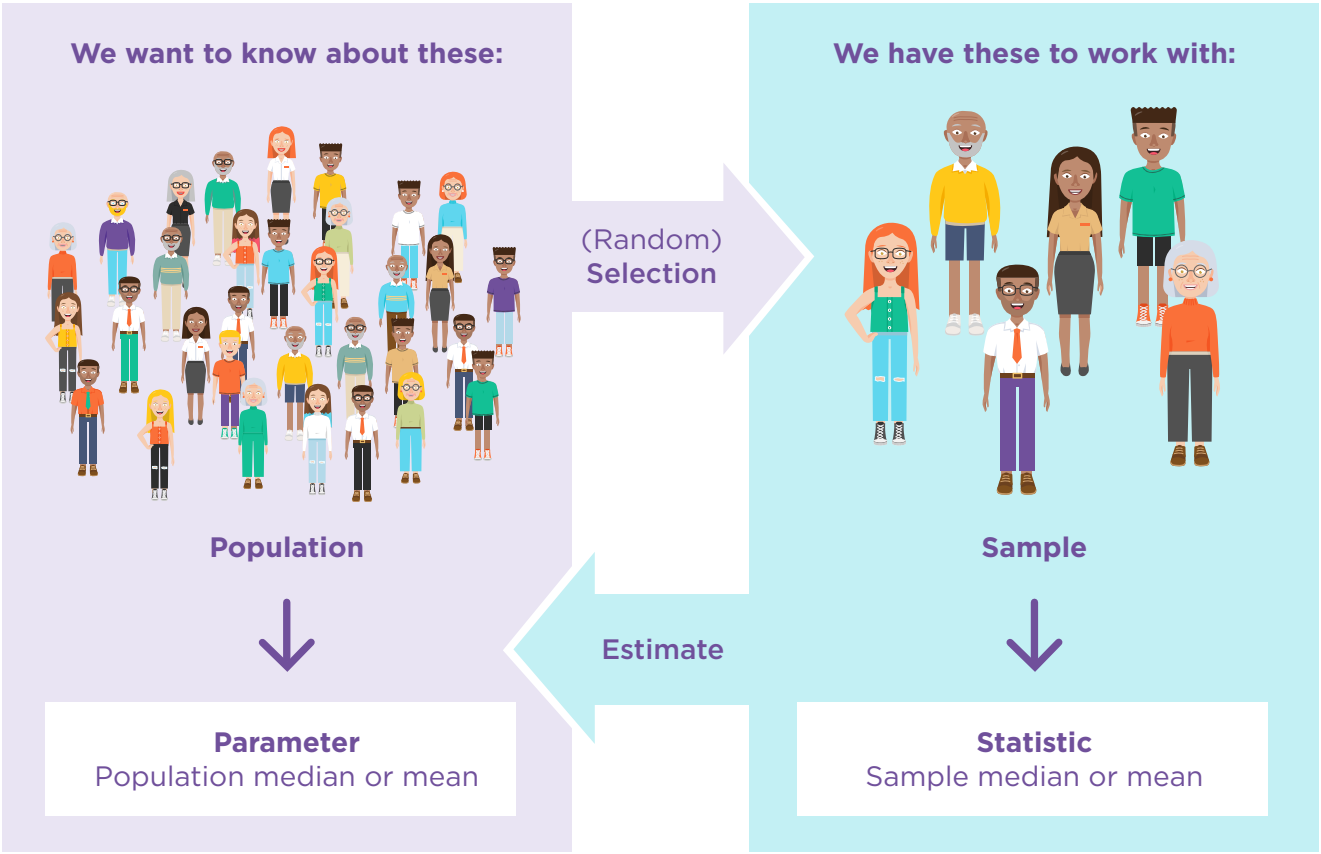
Both of these measures give us an idea about what a typical or "average" value is.

Measures that relate to a population are called **parameters**. Remember that although these measures exist, unless the population is small and easy to access, we usually can't find out what they are.

A **sample** is a group taken from a population. Measures that relate to a sample are called **statistics**.

If a **sample** is large enough and chosen carefully it should resemble the **population** it is taken from. As a result, we can use **statistics** from a **sample** to make an informed **estimate** of the **parameters** of a **population**.

Statistical Inference



Complete Topic 1 Activity 4
in the Student Practice Booklet

Median or mean?

The median and the mean are both averages. They give us an indication of what a typical value in a population might be.

When investigating variables such as income, savings, debt, or investment, the median is often a better average to use than the mean. This is because the median isn't influenced by unusually high or low values. For example, in any random sample of New Zealanders, there are likely to be some people who earn a great deal more than other people in the sample. If we calculate the mean of the sample, the very high incomes of a few people can have a big influence on the average, especially if the sample is small.

Complete Topic 1 Activity 5
in the Student Practice Booklet

In contrast, the median won't be affected by the very high levels of income of just a few individuals because it is simply the middle value. Therefore, the median is a better indicator of the typical income of the people in the sample.

This is why StatsNZ uses the median rather than the mean as an indicator of the gender pay gap. You can read more about their methodology in this **StatsNZ Measuring the gender pay** document.

Writing an investigative question

After we have established the population, sub-groups, variables and measures, we need to write an investigative question. The investigative question used in this module is:

the population *the measure* *the variable* *the two groups being compared*

In **2018**, was the **median KiwiSaver balance** of **females aged 55 to 64** lower than the **median KiwiSaver balance** for **males aged 55 to 64** in **Aotearoa New Zealand**?

Assessment tip:

Your investigative question needs to meet these requirements:

- The question needs to define which variable you are going to focus on, including units.
- The question needs to involve a comparison between two groups. The comparison needs to include a direction, for example, investigating whether the median of one group is bigger (or smaller) than the median of another group.
- Your question needs to state which parameter (median or mean) you are making an inference about.
- You need to fully define the population you are investigating, for example, by including age range, location, and year.

Complete Topic 1 Activity 6
in the Student Practice Booklet

Before moving on to Topic Two, check that you understand:

- [ ] How to define the purpose of an investigation
- [ ] What a population is
- [ ] The difference between categorical and numerical variables
- [ ] What the median and mean are
- [ ] The difference between a parameter and a statistic
- [ ] How to write an investigative question.

2

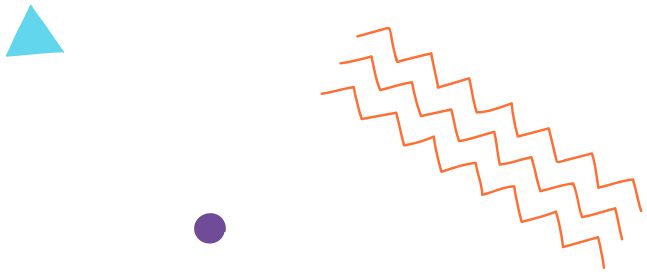
Topic Two:
Selecting a sample

Learning outcomes for Topic Two

- ☑ Understand the strengths and weaknesses of different sampling methods
- ☑ Know how to use software to select a sample.

Success criteria

- I can explain at least three sampling methods
- I can explain the strengths and weaknesses of at least three sampling methods
- I understand what a representative sample is.



We've already seen that in real life, it's often impossible to get information from everyone or about every item in a population.

As a result, sampling is part of almost every statistical investigation.

Assessment tip:

In your assessment, you'll be given a data set from a population. Instead of using the whole data set, the assessment activity will require you to take a sample and explain how you selected it.

Why not just use the whole data set?

Going through the process of selecting and analysing a sample gives you an opportunity to learn about the process and to discuss the strengths and weaknesses of that sampling method. Selecting a sample is also a requirement of the standard.

Ideally, a sample will be a good match for the population it comes from.

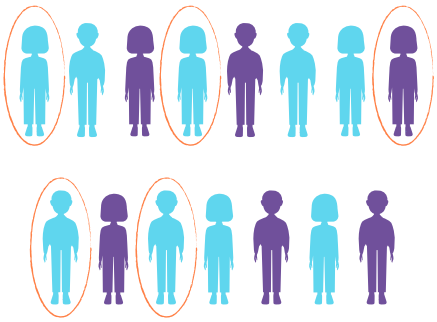
A sample is **unbiased** if everyone in the population has an equal chance of being selected to be in the sample.

A sample is **representative** if it accurately reflects the characteristics of the population. For example, if one third of a school are senior students, then a sample would be representative (in terms of juniors and seniors) if one third of the sample are seniors.

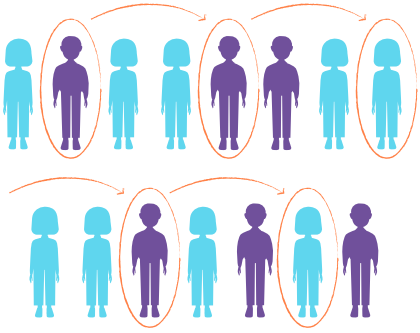
Sampling methods

The diagram below shows some different methods of selecting a sample (they are not the only ones). After the diagram, you'll find an explanation of these methods, along with their strengths and weaknesses.

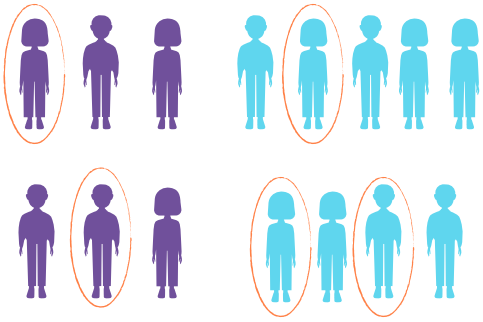
Simple random sample



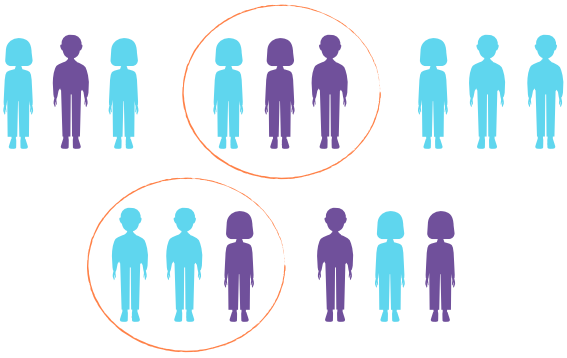
Systematic sample



Stratified sample



Cluster sample



Simple random sampling

Simple random sampling is an unbiased sampling method.

To conduct a simple random sample, you first need a list of everyone in your target population. Everyone on the list is given a number from 1 to the total number in the population. Then you can use a calculator or computer program to randomly select which numbers and values will be included.

Advantages of simple random sampling	Disadvantages of simple random sampling
Simple random sampling is unbiased – if you have a list of everyone in your population, everyone has an equal chance of being in the sample. If the sample is large, it should have similar characteristics to the population.	Simple random sampling can be time consuming and expensive. If your target population is large, it may not be possible to get a list of everyone in it. If the sample is small, it may not be representative. In other words, because the sample is selected randomly, certain groups in the population may be under or over represented.

Stratified sampling

Stratified sampling involves identifying sub-groups in the population, for example, the percentage of people who are employed full-time, employed part-time, or unemployed. The sample is selected in a way that means that each group makes up the same percentage of the sample as they do of the population. Importantly, once the sub-groups have been identified, the sample taken from each group is selected using random sampling methods.

Advantages of stratified sampling	Disadvantages of stratified sampling
Stratified sampling can make a sample more representative of the population than simple random sampling.	Stratified sampling can be time consuming.

Assessment tip:

Sorting your data into two categories (for example, males and females) doesn't make it a stratified sample, unless you also ensure that sub-groups within each category are represented in the same ways that they exist in the population. This introduces a degree of complexity not expected at this level.

Systematic sampling

Systematic sampling involves starting at a random place on a list and then choosing people at regular intervals, for example, every tenth or twentieth person.

Advantages of systematic sampling	Disadvantages of systematic sampling
Systematic sampling can be an easy way to test items being produced in a factory.	The sample may not be representative if there are recurring patterns in the list of the population or in the objects being produced.

Cluster sampling

Cluster sampling involves dividing the population into groups (clusters), for example, towns or regions of Aotearoa, and then selecting a random sample from one or more of these clusters rather than from the entire population.

Advantages of cluster sampling	Disadvantages of cluster sampling
Cluster sampling can save time and money because you are focusing on a smaller area.	The sample may not be representative. The clusters you use may not be a good match for the entire population.

You can learn more about sampling methods using this Statistical Learning Centre video Sampling: **Simple Random, convenience, systematic, cluster, stratified - Statistics Help**

Complete Topic 2 Activity 1
in the Student Practice Booklet

Selecting a sample using software

In your assessment, you will be using software to select a sample, for example, **NZGrapher, iNZight**, or **GenStat**. Check with your teacher which software you should use.

Assessment tip:

The software you will be using can select a random sample for you with a few clicks. This will ensure that your sample is unbiased.

Remember that an unbiased sample may not be representative of the population. For example, there may be a lower proportion of one sub group in the sample than there is in the population. This is something that you can discuss in your conclusion.

If you are using NZGrapher, watch this short video about how to take a simple random sample and a stratified sample: **NZ Grapher - sampling a data set**.

Note: If you haven't used NZGrapher before, this short video provides an **overview of the tool**. This NZGrapher webpage has information on **how to upload data**.

Complete Topic 2 Activity 2
in the Student Practice Booklet

Before moving on to Topic Three, check that you understand:

- [ ] What a representative sample is
- [ ] What an unbiased sample is
- [ ] The strengths and weaknesses of three sampling methods; simple random, stratified, and systematic
- [ ] How to select a sample using software such as NZGrapher.

3

Topic Three:
Displaying data and calculating statistics

Learning outcomes for Topic Three

- ✓ Understand how to describe what you can see in a dot plot
- ✓ Understand how to use software to create box plots and generate summary statistics
- ✓ Understand how to interpret box plots.

Success criteria

- I can describe the distributions of my samples
- I can explain what the different parts of a box plot represent
- I can use a statistical tool such as NZGrapher to create box plots and generate summary statistics.

Once you have selected a sample, the next step is to display the data you have collected and calculate statistics.

A common way to compare two samples is to display them as dot plots.

Discussing sample distributions

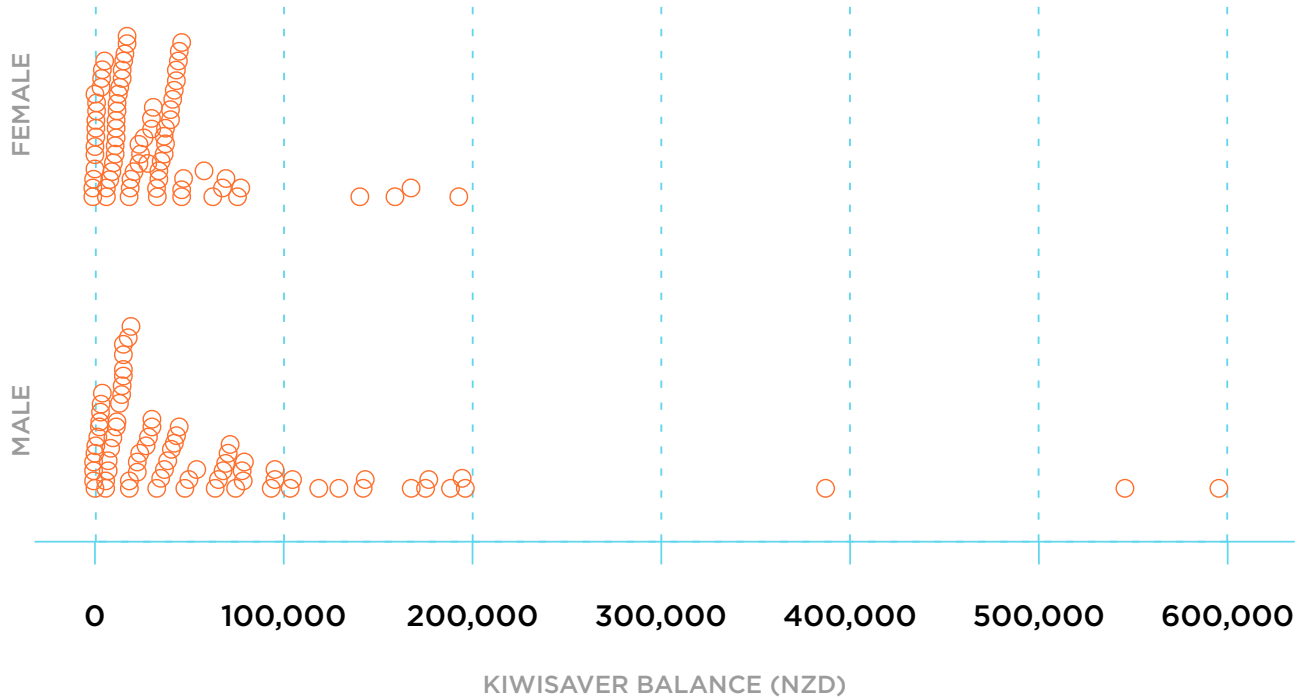
A dot plot helps you to see how the data in each sample is spread out, including where the middle of the data lies, what the most common value is, and whether there are any unusually large or small values. The way a data set is spread out is called its distribution.

Assessment tip:

In your assessment, it's important to discuss what you can see in the dot plot before moving on to specific statistics such as the median. Your discussion points need to be made in context. In other words, make sure you specifically refer to the two groups you have sampled and the variable you are comparing.

The dot plot below shows the KiwiSaver balances of a sample of 50 female New Zealanders aged 55 to 64 and a sample of male New Zealanders aged 55 to 64.

KiwiSaver balances of New Zealanders aged 55 to 64 (2018)



The dot plots reveal several features of the two sample distributions.

Centre

The middle of the male sample of KiwiSaver balances looks like it is higher than the middle of the female sample of KiwiSaver balances. We can see this because the female KiwiSaver balance data is more tightly clustered towards the lower end of the scale.

Symmetry

Neither of the two samples are symmetrical - the values in the top half of each sample are more spread out than the lower values. This makes the data skewed to the right. The sample of male KiwiSaver balances is more skewed to the right than the female sample of KiwiSaver balances.

Overlap

There is quite a lot of overlap between the two samples, particularly between $0 and about $80,000.

Outliers

Both samples have a few unusually large values. The most extreme value is in the male sample, with a KiwiSaver balance close to $600,000. The highest female KiwiSaver balance in the sample is around $200,000.

Complete Topic 3 Activity 1
in the Student Practice Booklet

Adding box plots

A visual inspection of the dot plots provides a useful overview of sample distributions. We can investigate the distributions of the samples in greater detail by adding a box plot.

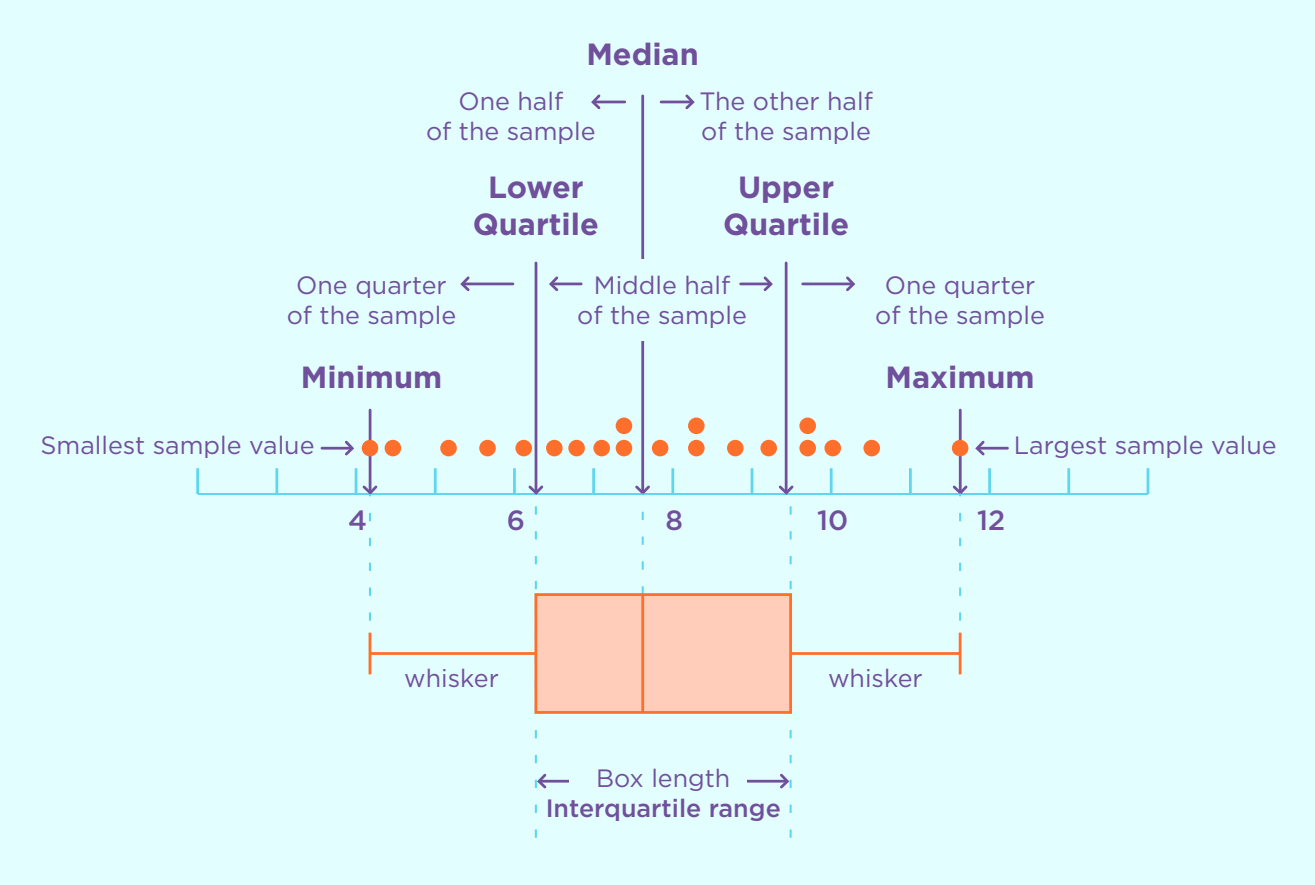
Assessment tip:

You can meet the requirements for an Achieved grade by discussing what you can see in the dot plots of your samples, making sure that you describe at least two features of the sample distributions. To achieve at Merit level or above, you need to create box plots and add summary statistics and then use these to support your discussion.

Box plots, sometimes called “box and whisker” plots, help us to explore the “shape” of the data and how it is spread out in more detail, for example, where the middle of the data is and whether the data is squeezed together or really spread out. Box plots also help to reveal outliers (unusually small or large values).

Interpreting a box plot

A box plot is made using five key statistics:



The left line of the box is the **lower quartile**. 25 percent of values lie below this line.

The line inside the box is the **median**. This is where the middle of the data is.

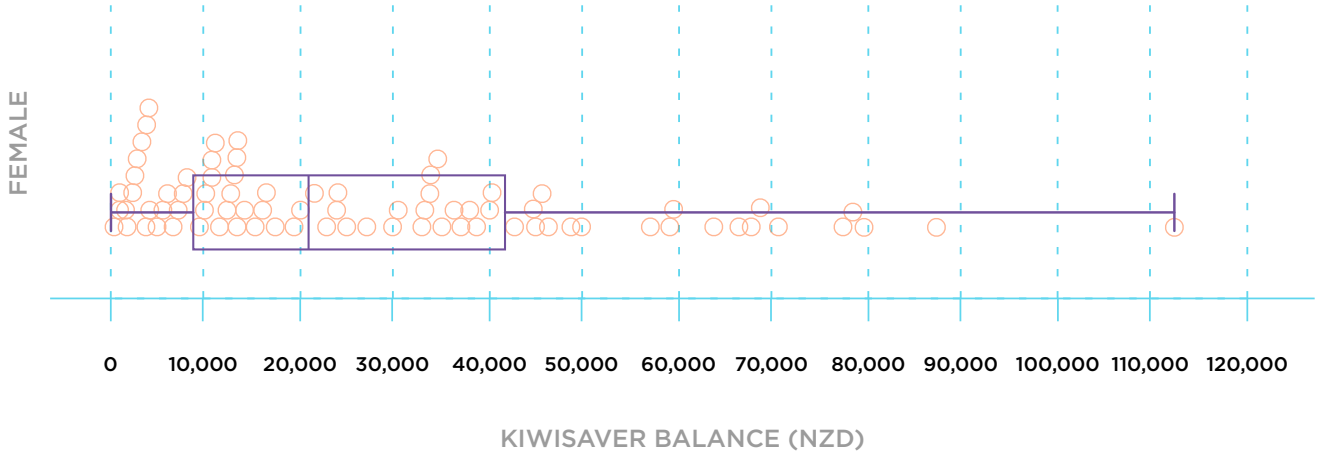
The right line of the box is the **upper quartile**. 25 percent of values lie above this line.

The box shows where the middle 50% of the data lies. When you analyse data using a box plot, the focus of much of your discussion will be on this middle half of the data.

The **minimum and maximum values** help us to see whether there are any unusually small or large values and provide information about the range of values in the population.

The box plot below has been constructed using CFFC Barometer data set information on KiwiSaver investments from a different sample of 80 females aged 55 to 64.

KiwiSaver balance of 55 to 64 year old women



Notice how each section of the box plot contains the same number of data values. For example, if you count the dots from the upper quartile up to the maximum value, you'll find that there are 20 dots. This is because there are 80 data values altogether and 80 ÷ 4 = 20. There are 20 dots in each of the other three sections of box plot.

The width of each section shows the spread of the data values within it. If a section is narrow, the data values within it lie close together; if the section is wide, the data values within it are spread out.

For example, in the graph above, the data values in the left half of the box (from the lower quartile to the median) are clustered together more tightly than the data values in the right half of the box (from the median to the upper quartile). As a result, the left-hand side of the box is narrower than the right-hand side.

Using software to create box plots

There are various online tools that you can use to create box plots. If you are using NZGrapher, this video explains **how to create a dot plot and a box plot**.

Complete Topic 3 Activity 2
in the Student Practice Booklet

Before moving on to Topic Four, check that you understand:

- [ ] How to use software to create a box plot
- [ ] How to add summary statistics to a box plot
- [ ] How to display two groups from the population.

4

Topic Four:
Comparing box plots and summary statistics

Learning outcomes for Topic Four

- ☑ Understand how to compare box plots
- ☑ Understand how to use summary statistics to support my statements.

Success criteria

- I can compare two box plots by referring to their central tendency, symmetry, spread, overlap, and unusual features
- I know how to remove data points from a data set.

In your report, you will need to use statistical language to compare how samples from the two groups you are investigating are spread out and where the centre of each sample lies.

Using PEEL to make statistical statements

PEEL is a framework you can use when you are writing about sample distributions.

P	**Point**	Make a statement about what you can see in the dot plot or box plot.
E	**Explanation**	Explain why you think this.
E	**Evidence**	Use numbers or facts to back up what you are saying.
L	**Link**	Link your statements to the population, the context, your research or to a prediction you made at the start of your investigation.

The PEEL framework has been modelled in the following examples of the features of box plots.

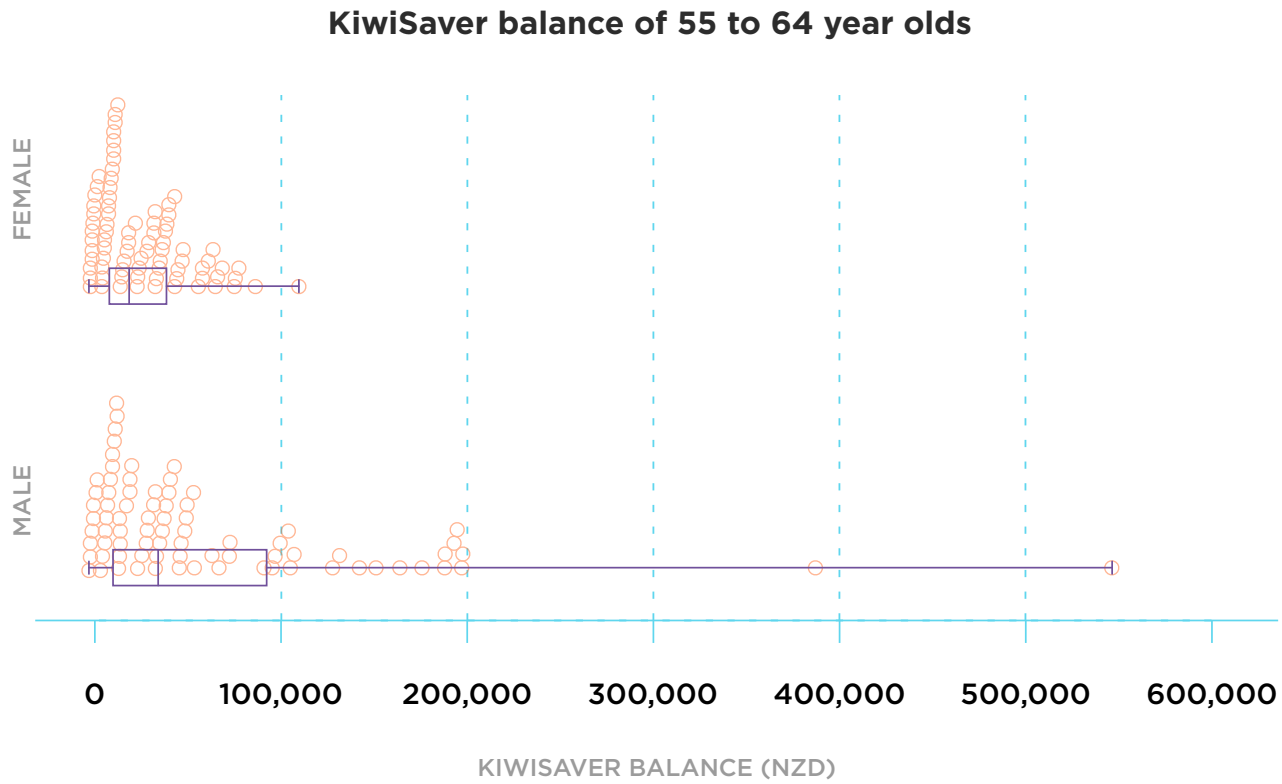
Making comparisons

Here are some features of box plots to refer to in your comparisons.

- Central tendency
- Shape
- Symmetry
- Spread
- Overlap
- Unusual features

Central tendency

Box plots make it easy to compare the central tendency (middle) of each data set because they show the median of each sample. You should compare the medians visually and then use the summary statistics to provide numbers to back up your statements.



Summary Statistics		
	Female	Male
Min	$157	$126
LQ	$8,851.5	$13,805
Med	$20,752.5	$36,522.5
Mean	$28,294	$66,133
UQ	$41,464	$95,121.5
Max	$110,989	$545,790
SD	$24,743	$87,215
Num	80	80

Point: In this sample, New Zealand males aged 55 to 64 had, on average, more money invested in KiwiSaver than the females in the sample.

Explanation: The sample median KiwiSaver for males aged 55 to 64 is higher than the sample median for females in the same age group.

Evidence: The summary statistics show that the median KiwiSaver investment for females in the sample was $20,752.50. The median KiwiSaver investment for males in the sample was $36,522.50.

Link: This lines up with the research I did on the gender investment gap, which suggests that males tend to be better set up financially for retirement than females.

Assessment tip:

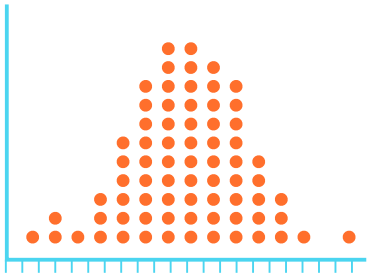
Remember that the median and mean of a sample are just estimates of population parameters. Just because the two samples have different medians, you can't assume that the medians of the populations are different without further analysis.

Shape

The shape of the dot plot provides useful information about how the data is spread out. We call this the distribution of a set of data.

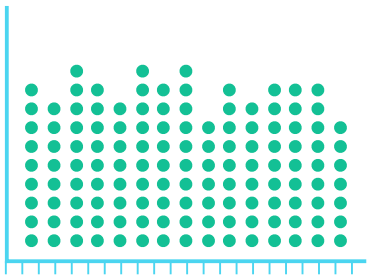
Bell-shaped

The data set below shows a bell-shaped distribution. Most of the data values are clustered in the middle of the data set. The mode (most common value) is around the middle of the data.



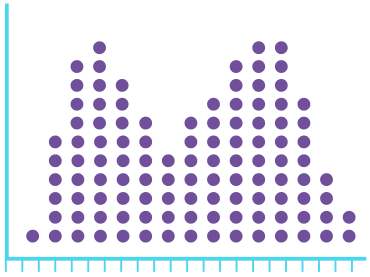
Uniform

The data set below shows a uniform distribution. A uniform distribution looks a bit like a rectangle.



Bi-modal

This data set looks like it might have two distinct groups in it. If there are two highest point, each with data values clustered around them, we say that the data is bimodal. Bimodal data can suggest that there are subgroups in the data that are influencing the values.



Symmetry

A box plot is symmetric if the left side of the graph is a mirror image of the right side of the graph and the median lies half-way between the lower quartile and the upper quartile. If a box plot is symmetric, it means that the data is spread out evenly.

Symmetric

If the graph is not symmetric, it means that some parts of the data are clustered together more closely than other parts, making the data skewed. If the data is stretched out on the left-side, we say that it is skewed to the left.

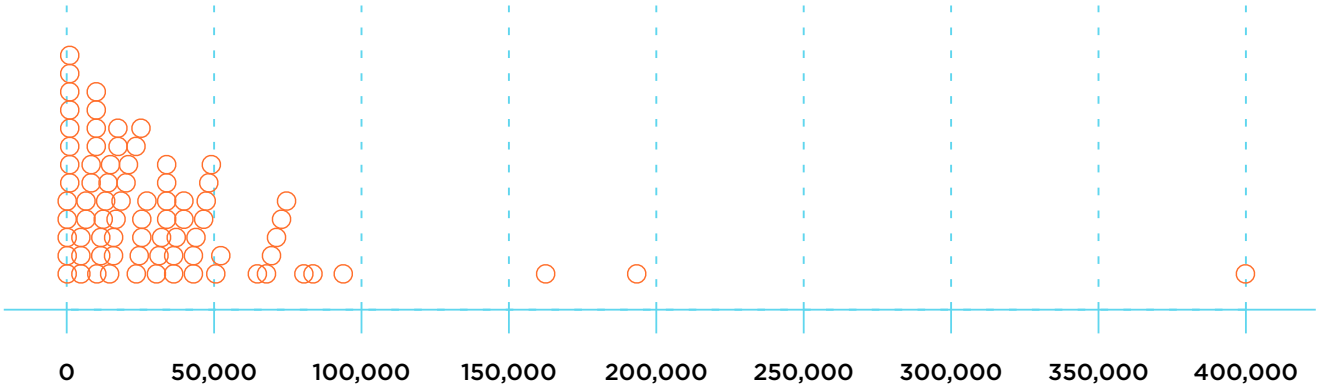
Skewed to the left

If the data is stretched out on the right side, we say that it is skewed to the right.

Skewed to the right

This graph shows the KiwiSaver balances of a random sample of 80 New Zealand females ages 55 to 64.

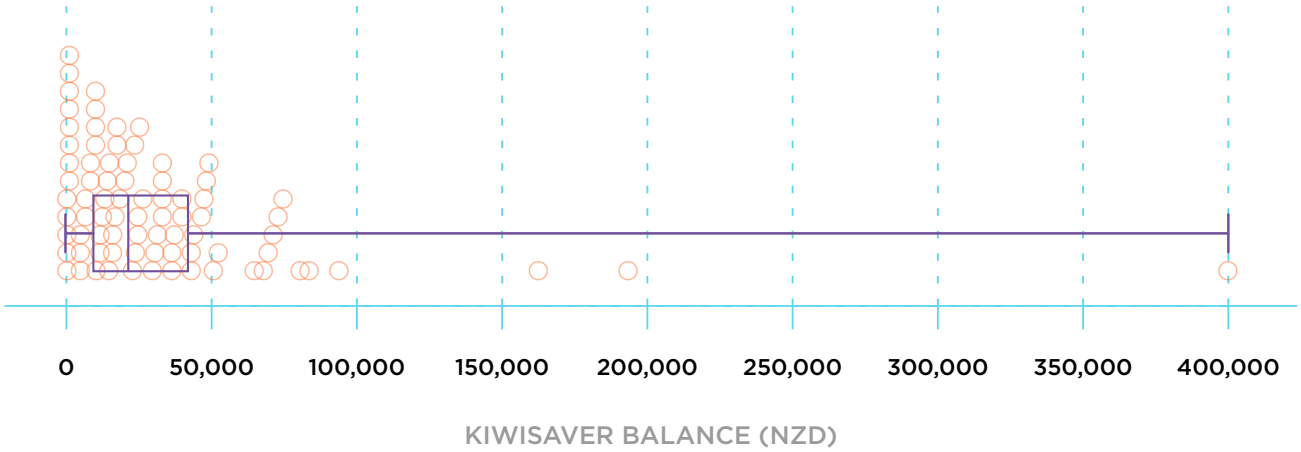
KiwiSaver balances of New Zealander women aged 55 to 64 (2018)



Looking at the dot plot, we can see that the data is skewed to the right. There is also at least one outlier (the values close to $400,000).

Now let's add a box and whisker plot and summary statistics:

KiwiSaver balances of New Zealander women aged 55 to 64 (2018)



Summary Statistics	
	Female
Min	$559
LQ	$8,161
Med	$19,678.5
Mean	$34,833
UQ	$43,850
Max	$399,064
SD	$52,633
Num	80

Point: The box plot confirms that the KiwiSaver balance of females aged 55 to 64 is skewed to the right.

Explain: This means that there are lots of values in the lower half of the data that are quite close together and more variation in the top half over the data, especially in the top 25 percent.

Evidence: For example, the lower half of the box has a width of over $11,000 (median – lower quartile = $11,517), while the upper half of the box has a width of over $24,000 (upper quartile – median = $24,172).

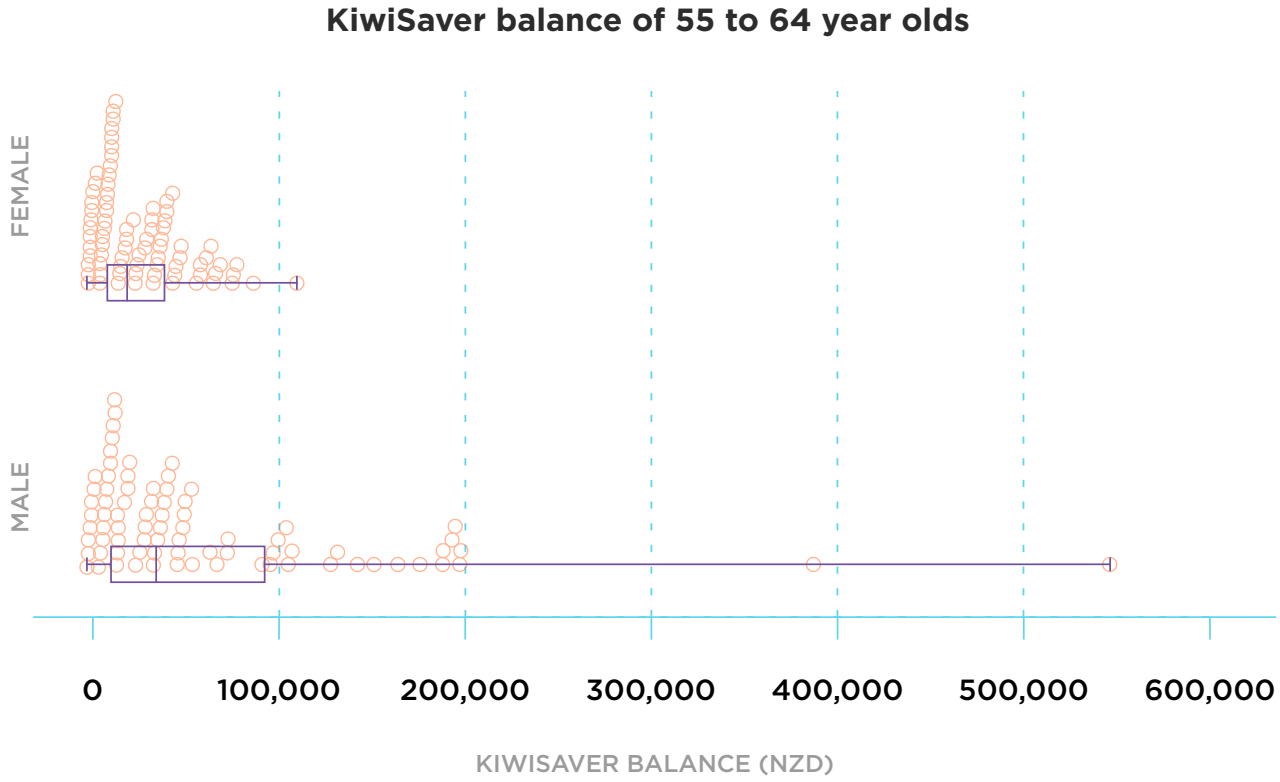
Link: This lack of symmetry reflects that there is a high level of inequality in Aotearoa New Zealand. The lower 25 percent of females in the sample have KiwiSaver balances that won't stretch very far once they reach retirement, while the top 25 percent of females in the sample have at least four times as much money as the lower quartile KiwiSaver balance.

Take a moment to compare the median and the mean of this sample. The median is $19,678 and the mean is $34,833. The median and the mean are both "measures of central tendency" – they provide information about where the centre of the data lies. When the data is skewed, in this case to the right, the mean gets pulled in the direction of the skew.

The median is not influenced by unusually large or small values because it's simply the middle value. This makes the median a better indication of a typical KiwiSaver balance for the females in this sample than the mean.

Spread

You can compare the spread of the two samples by comparing the widths of different sections of each box plot. The interquartile range is defined as the width of the box. You can calculate it by subtracting the lower quartile from the upper quartile.



Summary Statistics		
	Female	**Male**
Min	$157	$126
LQ	$8,851.5	$13,805
Med	$20,752.5	$36,522.5
Mean	$28,294	$66,133
UQ	$41,464	$95,121.5
Max	$110,989	$545,790
SD	$24,743	$87,215
Num	80	80

Point: In this sample, there is more variation in the male KiwiSaver data than in the female KiwiSaver data.

Explain: Looking at the middle 50 percent of each sample, the male KiwiSaver balances are much more spread out than those of females.

Evidence: For this sample, the interquartile range of the female KiwiSaver investments is $41,464 – $8,851.50 = $32,612.50. The interquartile range of the male KiwiSaver investments is $95,121.50 – 13,805 = $81,316.50.

Link: This could reflect the fact that females have less career opportunities (and therefore earning capability) than males, which is why there is less variation in their KiwiSaver balances.

Unusual features

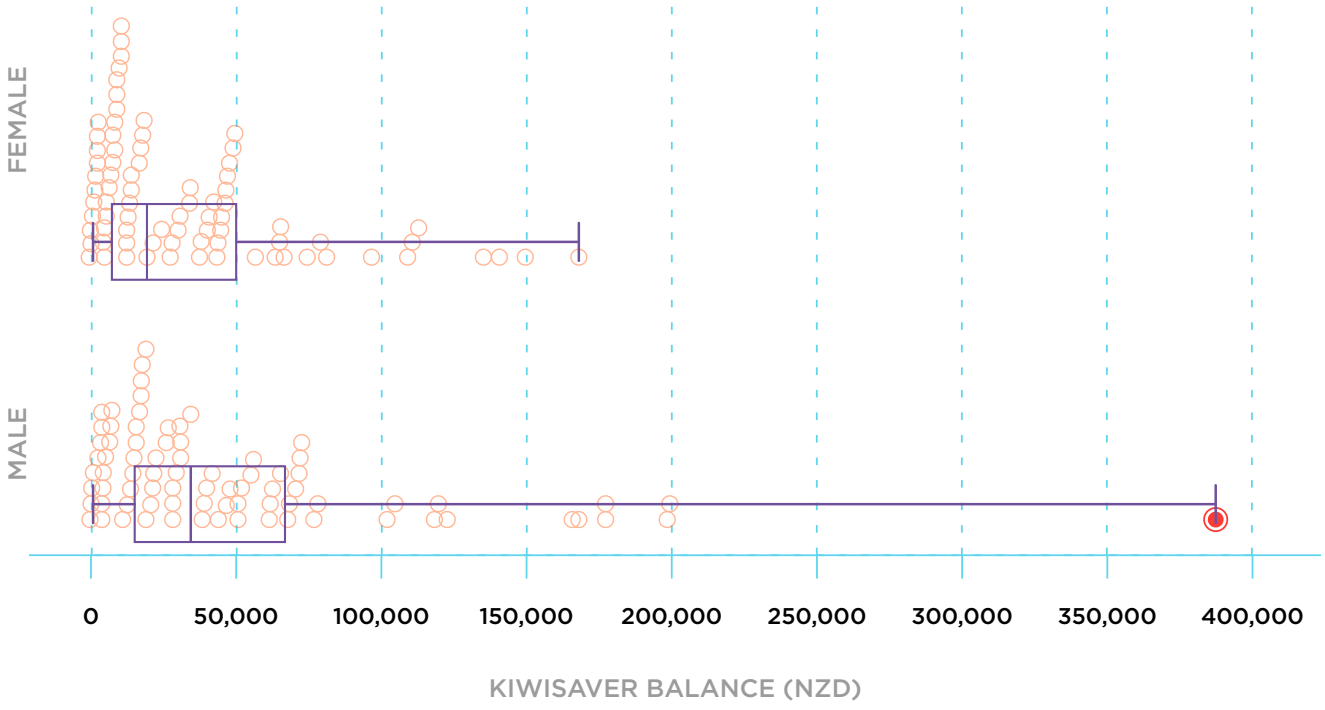
The end points of the box-and whisker graph show where the minimum and maximum values are. In general, you should focus more on where the middle 50 percent of the data lies. However, in your report, you should refer to these endpoints if they show that one or two values are unusually high or low.

Generally, if a data value lies more than one-and-a-half box-widths below the lower quartile or above the upper quartile of the graph, it is called an outlier.

Outliers and extreme outliers are sometimes shown as individual points lying away from the majority of values on the graph. If a data value lies more than three box-widths below the lower or above the upper quartile of the graph, it is called an extreme outlier.

Outliers could be an error or they could show that something unusual has happened.

KiwiSaver balances of New Zealanders aged 55 to 64 (2018)



Summary Statistics		
	Female	Male
Min	$447	$532
LQ	$8,000.5	$14,552
Med	$19,428.5	$30,690
Mean	$35,214	$51,809
UQ	$47,163.5	$67,986
Max	$168,162	$387,214
SD	$38,232	$62,005
Num	80	79

Point: In the box plots above, there is at least one obvious outlier (circled in red).

Explain: This point can be considered an extreme outlier because it lies more than three box-widths above the middle percent of the data.

Evidence: The maximum male KiwiSaver balance is $387,214. The distance between this value and second highest value is similar to the total range of all the other data values.

Link: Although this data value lies a long way from the box part of each graph, it is probably not an error. Income inequality is a significant issue in Aotearoa New Zealand, it's not surprising that some people have a KiwiSaver balance that is much higher than most of the other people in the population. The graph simply reflects this reality.

Removing outliers

It can be very tempting to remove an outlier from your sample, but this should only ever be done for a good reason. Generally, you should only remove an outlier if it is obvious that it is a mistake (for example, someone's height is recorded as 15cm).

Assessment tip:

You can remove data values on NZ Grapher by:

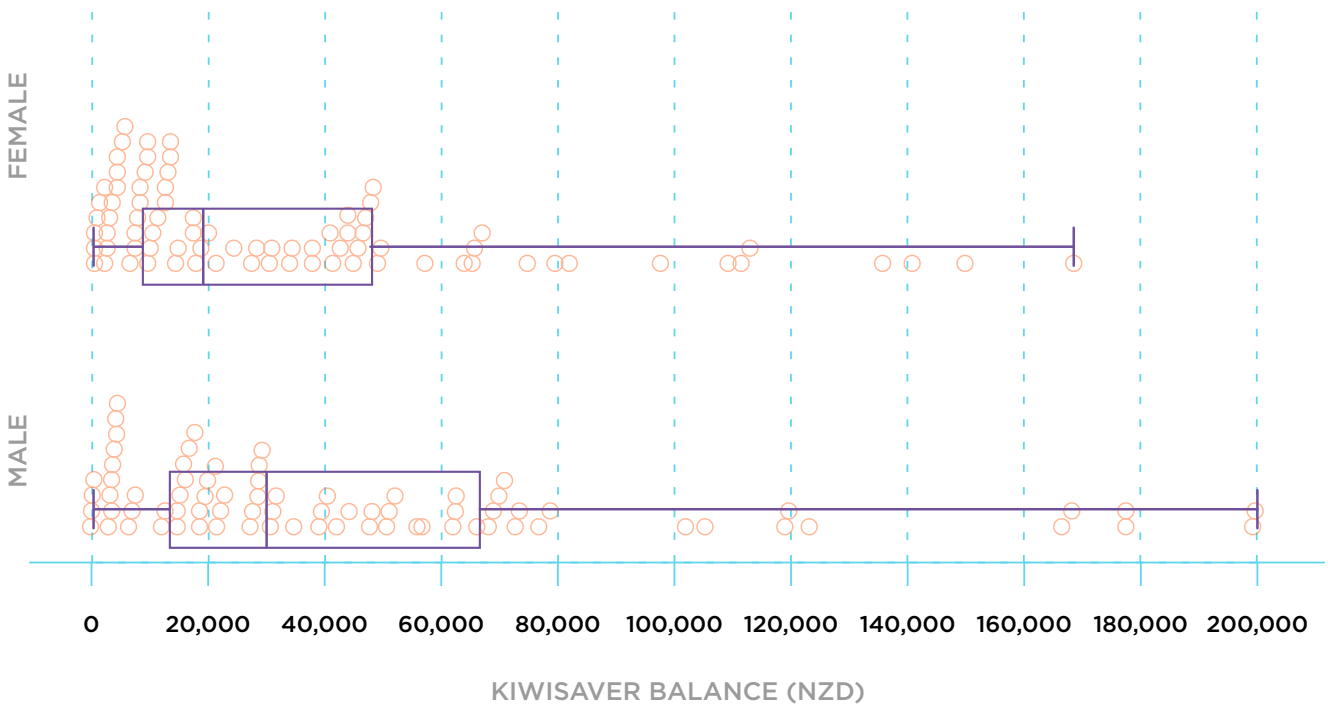
- hovering the cursor over the outliers to find out their Point ID
- selecting Row and then Remove Specific Row from the drop down menu that appears.

Note that after you remove one data value, the Point ID of other outliers might change, so use this option carefully. You can avoid this by deleting the rows of data points with higher Point IDs first.

If you remove an outlier, make sure that you explain why in your report.

In this sample, it's unlikely that the outlier is a mistake, but because it is so much higher than the rest of the data values, it makes it hard to see how the rest of the data is spread out. We can remove it to see what happens, and to make it easier to compare the rest of the data, but it definitely needs to be included in our discussion.

KiwiSaver balances of New Zealanders aged 55 to 64 (2018)



Summary Statistics		
	Female	**Male**
Min	$447	$532
LQ	$8,000.5	$14,552
Med	$19,428.5	$29,898.5
Mean	$35,214	$47,508
UQ	$47,163.5	$65,933
Max	$168,162	$199,004
SD	$38,232	$49,140
Num	80	78

Take a moment to notice how removing the two outliers impacted on the medians and means of the sample of male KiwiSaver balances:

Male data with outlier	Male data without outlier
Median: $30,690	Median: $29,898.50
Mean: $51,809	Mean: $47,508

Removing the largest value from the male sample has had a much bigger impact on the mean than the median of each group. This is because the median is not influenced by extremely large values.

Removing this outlier does make it easier to focus on the middle 50 percent of the data. We might decide to continue our investigation without the outlier, but the outlier definitely provided useful information about the level of inequality in Aotearoa New Zealand and should be mentioned in our analysis and conclusion. The outlier suggests that some people enter retirement with really large sums of money. These people are likely to have a range of other assets as well.

Taking another random sample would help us to see how unusual the outlier really is.

Complete Topic 4 Activity 1
in the Student Practice Booklet

Before moving on to Topic Five, check that you understand:

- [ ] How to compare two box plots by referring to their central tendency, symmetry, spread, overlap, and unusual features.
- [ ] How to remove data points from NZGrapher.

5

Topic Five: Confidence intervals

Learning outcomes for Topic Five

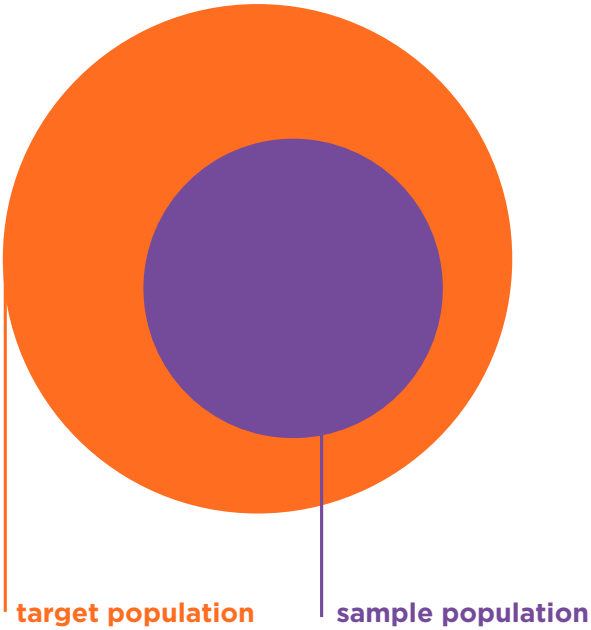
- ☑ Understand what sampling variability is
- ☑ Understand what confidence intervals are and how to interpret them

Success criteria

- I can explain causes of sampling variability
- I can use a formula or a statistical tool such as NZGrapher to construct a confidence interval
- I can explain two factors that influence the width of a confidence interval
- I can interpret a confidence interval and use it to make an inference about the population.

Understanding sampling variability

Remember that a sample is only a part of the population, so we can never guarantee that it is a perfect match for the population; if we take a different sample, we will almost certainly get a different result. This is called sampling variability.



If we know that two samples taken from the same population will have different sample statistics and box plots, how can we be sure that our estimates of a population parameter is accurate? The short answer is that we can never be 100 percent certain that we have accurately estimated a population parameter.

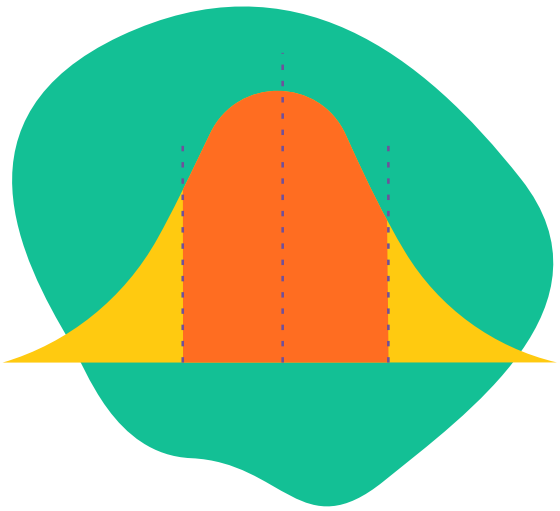
One way to get around this is to provide an interval of values that we think is likely to include the true median or mean rather than just stating a single value or "point estimate".

Complete Topic 5 Activity 1
in the Student Practice Booklet

Confidence intervals

A confidence interval is a set of values based on a sample statistic that is likely to include the true population median (or mean).

Because confidence intervals provide a sensible range of possible values for the true population parameter, we can use them to make statements about whether our investigation supports our hypothesis about a potential difference between two medians (or means). They are, however, not 100 percent foolproof.

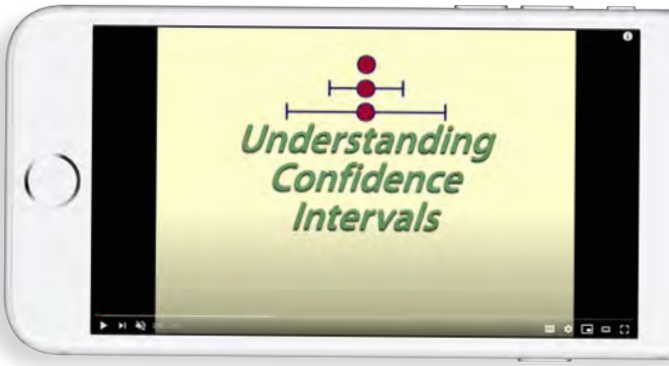


The width of the interval reflects how accurate the estimate is likely to be. A narrow interval suggests that we are very confident that our estimate is accurate. A wide interval suggests that there is a lot of uncertainty about the accuracy of our estimate.

The larger our sample, the less uncertainty there will be about our findings.

Watch **Dr Nic's Maths and Stats video about confidence intervals**.

As discussed in the video, the width of a confidence interval is determined by the amount of variation in the population (less variation = narrower confidence interval) and the number of data values in the sample (larger sample = narrower confidence interval).



Constructing informal confidence intervals

There are different ways to construct an informal confidence interval.

Using a formula

An informal method involves adding and subtracting a fixed amount from the median. The amount that is added and subtracted is:

$$median \pm 1.5 \times \frac{IQR}{\sqrt{n}}$$

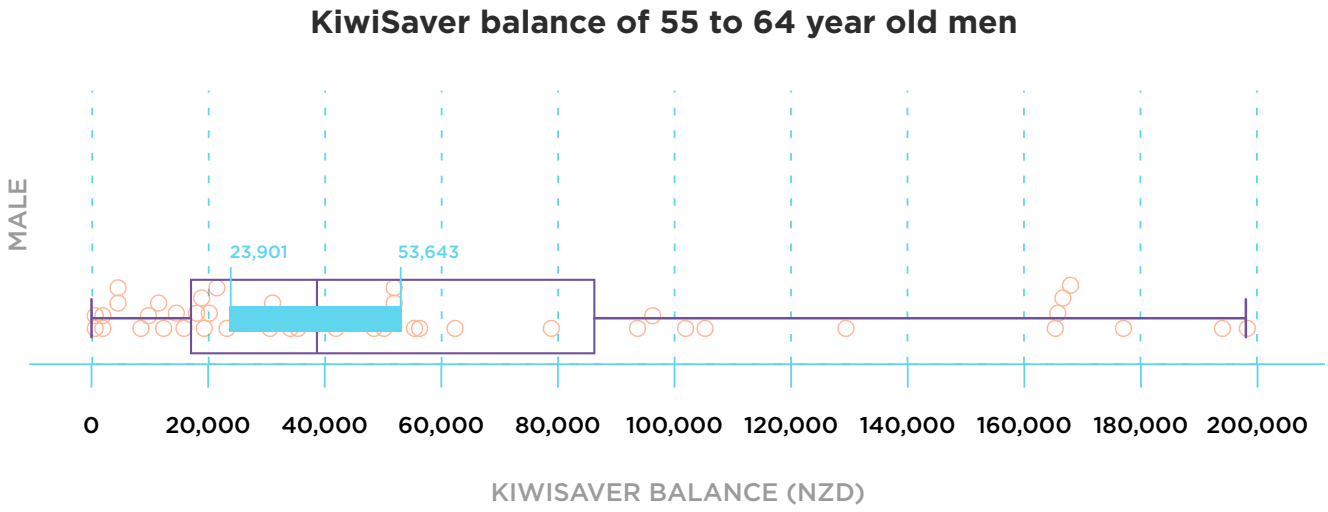
IQR is the interquartile range (the upper quartile of the sample minus the lower quartile of the sample) and n is the number of data values in the sample.

90 percent of confidence intervals constructed in this way will contain the actual value of the population median. (Note that this means that 10 percent of confidence intervals constructed using this method won't contain the population median, so we haven't removed all of the uncertainty of our estimate simply by providing an interval rather than a single value.)

Complete Topic 5 Activity 2
in the Student Practice Booklet

Displaying informal confidence intervals on a box plot

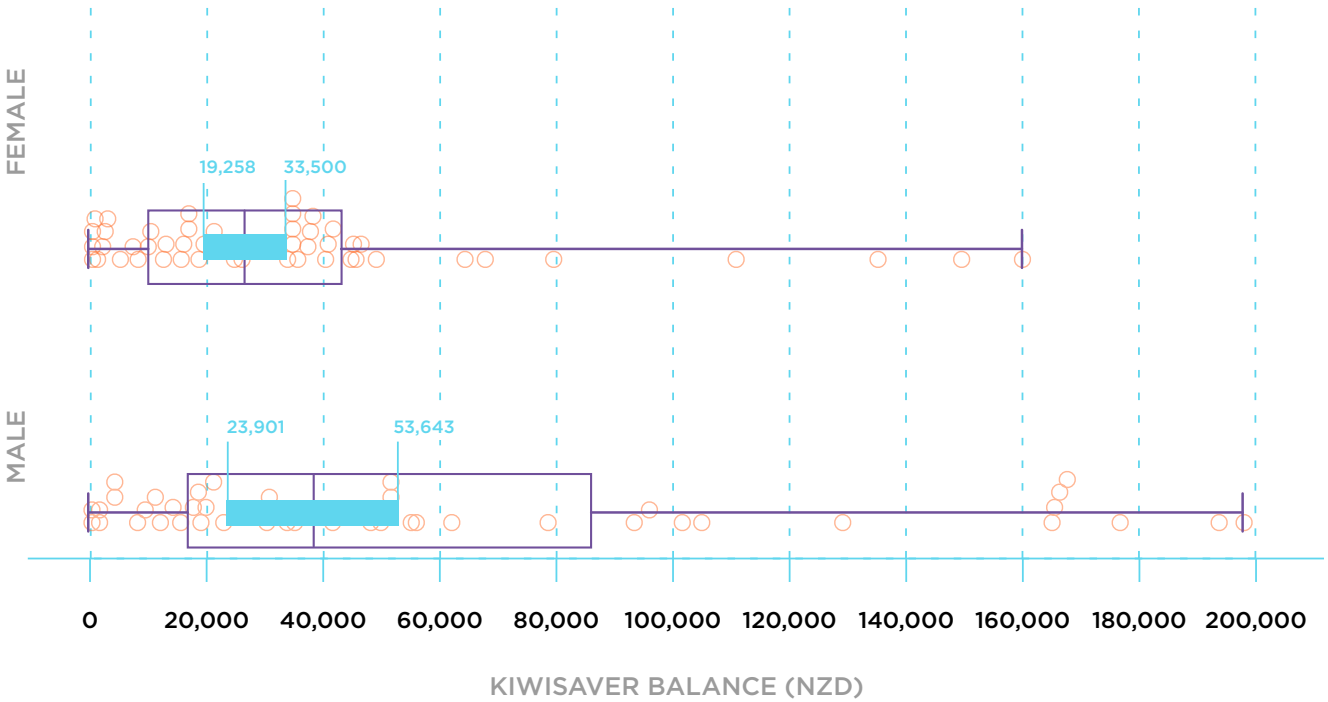
Confidence intervals can be drawn onto a box plot as horizontal line:



Notice how the median is the exact midpoint of the confidence interval. This is because the same amount has been added and subtracted from the median ($1.5 \times \frac{IQR}{\sqrt{n}}$).

Using NZGrapher to create an informal confidence interval

If you're using NZGrapher, this video shows you how to **add a confidence interval** to your graph and find the lower and upper limits of confidence interval.



Summary Statistics		
	Female	Male
Min	$157	$608
LQ	$10,109.5	$16,956.5
Med	$26,379	$38,772
Mean	$35,189	$57,664
UQ	$43,339	$86,354
Max	$160,200	$198,504
SD	$36,907	$57,450
Num	49	49

Using this sample, the informal confidence intervals for the KiwiSaver balances for males and females are:

Group	Lower limit	Upper limit
Females	$19,258	$33,500
Males	$23,901	$53,643

Notice that the confidence interval for the male data is much wider than the confidence interval for the female data. This is because the sample data for the males is more spread out than the sample data for the females. This suggests that there is more variation in male KiwiSaver balances in the population, which means that a second sample might produce quite different results. This level of uncertainty is reflected in the greater width of the male KiwiSaver balance confidence interval.

Interpreting confidence intervals

The wording used to interpret a confidence interval is really specific. Based on the confidence intervals we have constructed we can say:

> *"We can be reasonably sure that back in the population of New Zealanders aged 55 to 64 in 2018, the median KiwiSaver balance of females lay somewhere between $19,258 and $33,500 and that the median KiwiSaver balance of males aged 55 to 64 lay somewhere between $23,901 and $53,643."*

Notice that the statement doesn't imply 100 percent confidence in our results. This is because we don't know whether our sample was a good match for the population and we know that if we took another sample, the results could be different.

Assessment tip:

Remember that in your assessment it is important to find out the units for any data you work with, and to use these units whenever you refer to an amount or value. In this instance, the unit is New Zealand dollars, but just using a $ sign when you write the confidence interval is fine. It's obvious from the context that we're talking about New Zealand dollars.

Using confidence intervals to draw a conclusion

We've come up with an interval of values that we are reasonably confident contain the true median. How do we use these to draw a conclusion?

Firstly, let's return to our investigative question:

> In 2018, was the median KiwiSaver balance of females aged 55 to 64 lower than the median KiwiSaver balance for males aged 55 to 64 in Aotearoa New Zealand?

We need to make a call about whether the population median KiwiSaver balance of females is less than that of males in the age group we are investigating. Our answer will be based on the two confidence intervals we have constructed.

Evidence that there is a difference between the medians of two groups

We can say that there is sufficient evidence to suggest that there is a difference between the medians of the two groups in the population if the confidence intervals for the two medians don't overlap (in other words, the upper limit of one group is lower than the lower limit of the other group).

Insufficient evidence that there is a difference

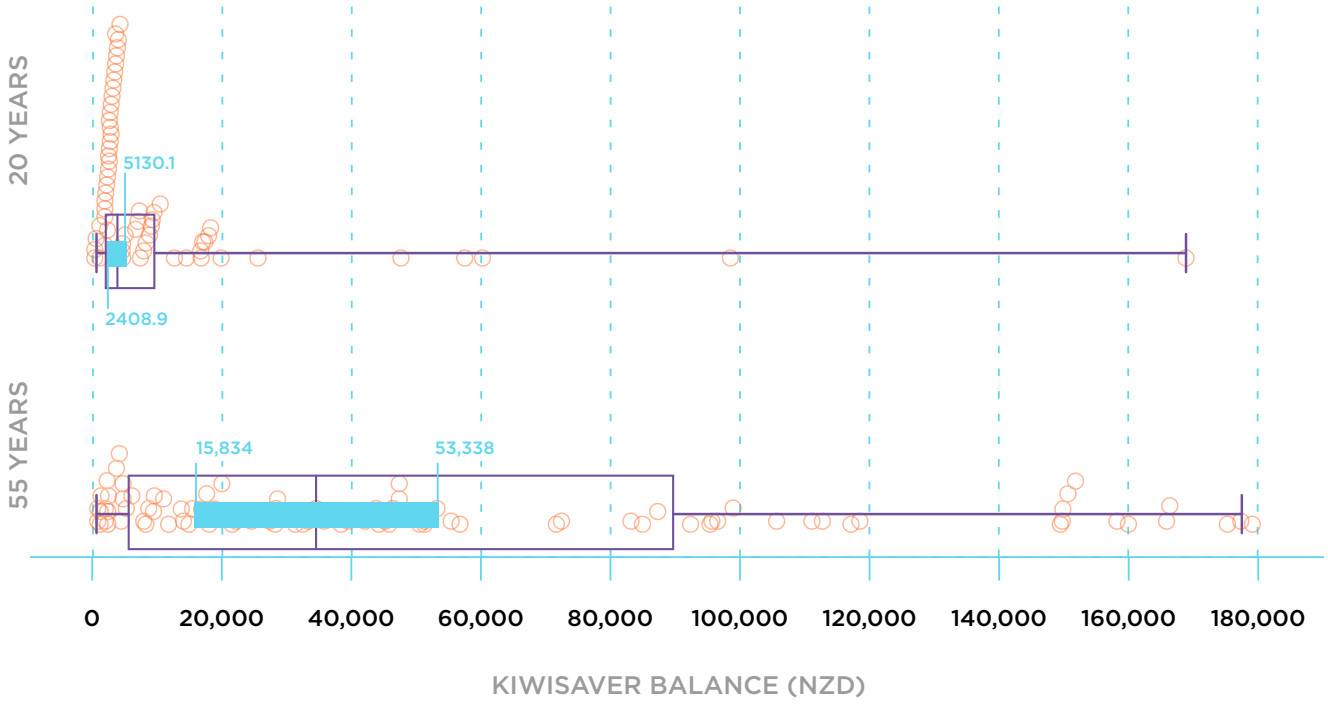
If the informal confidence intervals for the two median overlap, then we don't have enough evidence to suggest that the medians of the two groups in the population are different.

In our example, the two confidence intervals overlap. Therefore, based on this sample, we cannot say that there was a difference between the median KiwiSaver balance of females and males in the 55 to 64 age group in Aotearoa New Zealand in 2018.

Assessment tip:

In your assessment, you need to base your conclusion about the population medians on your confidence intervals, not the difference between the two sample medians.

Here is an example of a comparison of two groups where the confidence intervals do not overlap. The graph shows the distribution of samples of KiwiSaver balances of people aged 20 and 55.



In this example, the informal confidence intervals do not overlap, and based on this sample we could be reasonably sure that back in the population the median KiwiSaver balance of people aged 20 is less than the median KiwiSaver balance of people aged 55.

Complete Topic 5 Activity 3
in the Student Practice Booklet

Discussing the variability of estimates

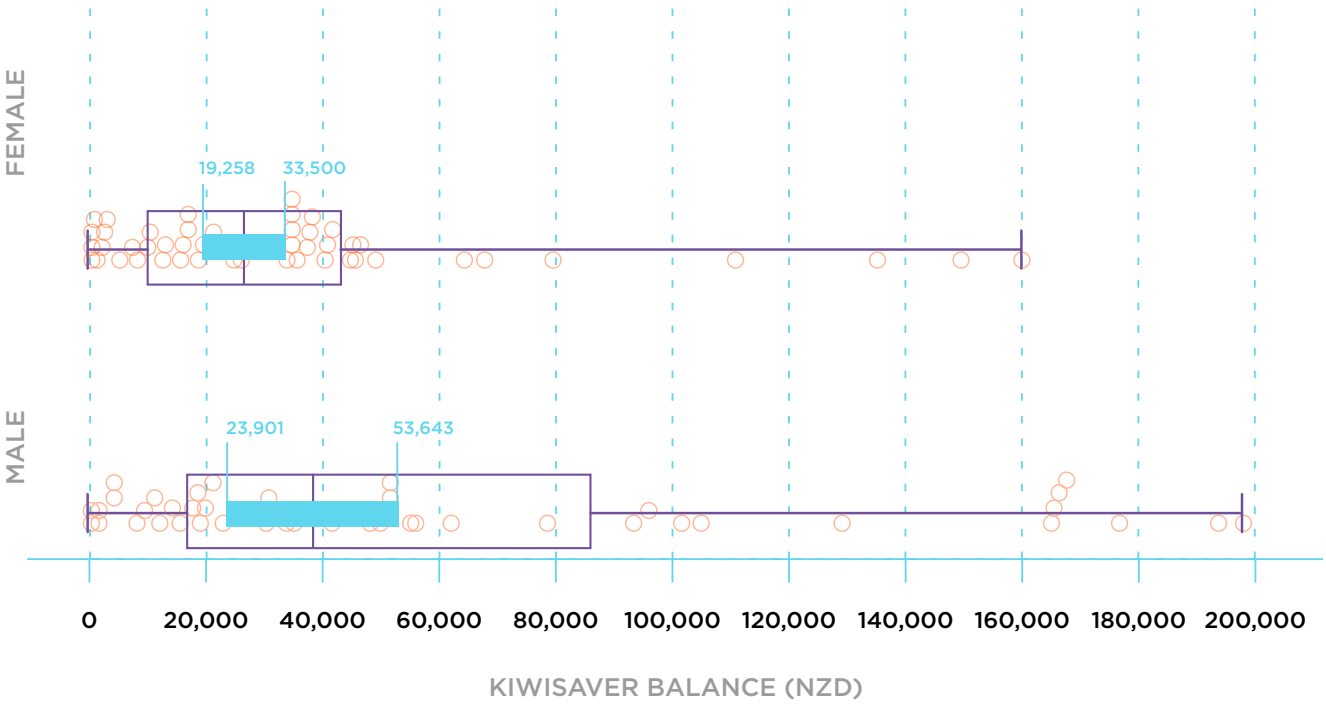
Whichever conclusion we draw, we need to make it clear that our conclusion is based on a single sample, and that if we were to take another sample and repeat the process, we might come up with a different conclusion. This is because no two samples are ever the same. It is possible that our randomly selected sample contained an unusually high number of small or large values.

Assessment tip:

Discussing sampling variation, including the variability of samples, is an essential part of your assessment. If you complete the process of making a statistical inference, including drawing a correct conclusion based on your confidence intervals, but don't mention sampling variability, you won't pass.

Although we can't completely remove a degree of uncertainty about our conclusion, the size of our sample plays a role in terms of how confident we are that our sample is a good match for the population. A large sample means that there is less uncertainty that our sample represents the population.

KiwiSaver balance of 55 to 64 year olds



Summary Statistics		
	Female	Male
Min	$157	$608
LQ	$10,109.5	$16,956.5
Med	$26,379	$38,772
Mean	$35,189	$57,664
UQ	$43,339	$86,354
Max	$160,200	$198,504
SD	$36,907	$57,450
Num	49	49

In the KiwiSaver balance data set, we can see that there is a lot of variation in the male sample. This variation is reflected in the width of the confidence interval. This variation does make it harder to be sure that a second sample would provide a similar result.

If the overlap between the two confidence intervals had been a lot smaller, then you would have additional grounds for stating that you're not very confident in your inference. A minor overlap means that an inference from a different sample might result in a different conclusion due to sampling variability.

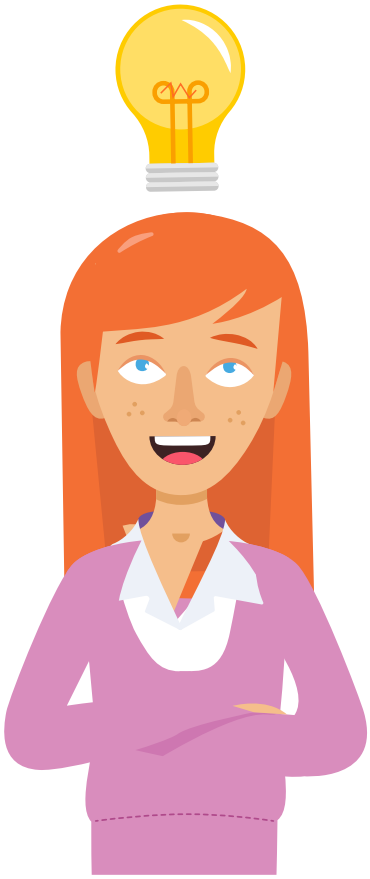
Evaluating the process

The final stage of the statistical enquiry process involves evaluating the process you went through to make an inference and use it to draw a conclusion. In this section of your report, you can discuss assumptions you have made, the limitations of the process, or improvements that could be made.

There is no set formula for how to evaluate your own conclusions. Try to think about sub-groups that weren't acknowledged in the sampling process or areas of further investigation.

For example, our exploration of KiwiSaver balances only compared males and females because we only had access to information about these two groups. As a result, gender diverse New Zealanders were excluded from our study.

We used simple random sampling to select a sample from each group we were investigating (male and female New Zealanders aged 54 to 65). However, we know that Māori and Pasifika females tend to provide a lot more support for their extended whānau and communities.

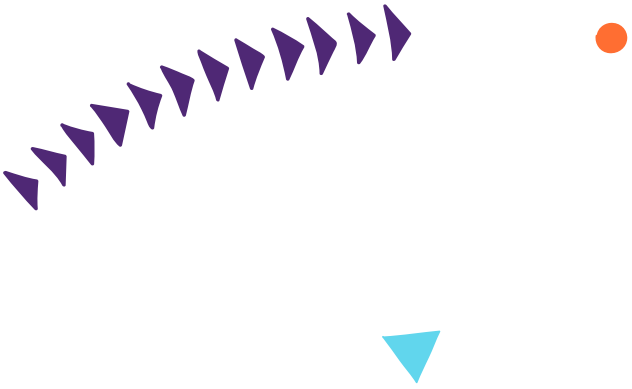


This is unpaid work, so Māori and Pasifika females may be more likely to be impacted by a gender investment gap than Pākehā New Zealanders. A limitation of our investigation was that we did not explore this group separately or use stratified sampling to make sure that they were adequately represented in our sample.

Another area of investigation would be to explore a younger age group, for example, males and females aged 35 to 54. This is an age group in which females are more likely to have family commitments that impact on their earning opportunities.

Before moving on to Topic Six, check that you understand:

- [ ] Why no two samples will ever be identical
- [ ] How to use a formula or NZGrapher to construct a confidence interval
- [ ] Factors that influence the width of a confidence interval
- [ ] How to interpret a confidence interval and use it to make an inference about the population
- [ ] How to evaluate your conclusion.



6

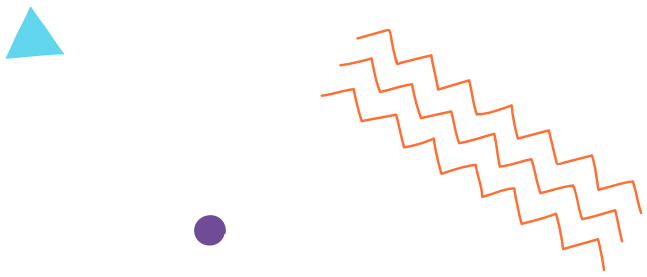
Topic Six:
Writing your report

Learning outcomes for Topic Six

- ☑ Understand how to structure your report
- ☑ The content you need to include in your report.

Success criteria

- I can write a report based on a statistical inference.

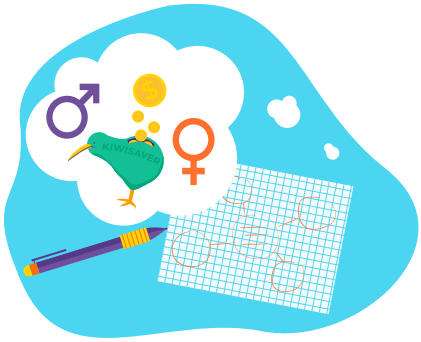


This section of the module explains how to write your report.

It includes important information about wording, particularly when it comes to interpreting confidence intervals, and on the information you need to provide for Achieved, Merit, and Excellence grades.

The report framework is structured around the statistical enquiry cycle.

Problem and plan



Give your report a title that shows what it is about, for example:

> *Investigation into the gender investment gap in Aotearoa New Zealand.*

Explain the purpose of your investigation – what are you going to look into and why is this important or useful.

Write an investigative question that involves comparing two medians. Make sure that it has a clear link to the purpose of your investigation, for example:

> *In 2018, was the median KiwiSaver balance of females aged 55 to 64 lower than the median KiwiSaver balance for males aged 55 to 64 in Aotearoa New Zealand?*

Make sure that your investigative question includes:

- The variable you are going to focus on, including units
- A comparison of two groups; the comparison needs to include a direction, for example, investigating whether the median of one group is bigger (or smaller) than the median of another group
- Which parameter (median or mean) you are making an inference about.

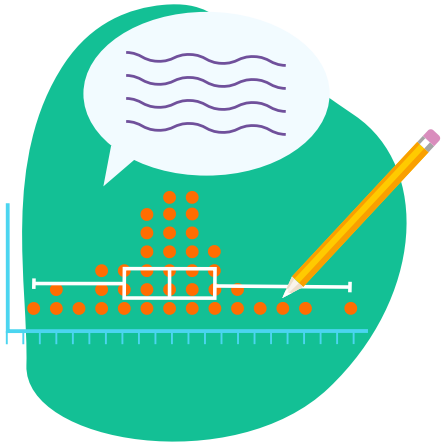
You need to fully define the population you are investigating, for example, by including age-range, location, and year.

Explain where the data you are using is from.

Identify which sampling method you will use, including the number of data values you will use.

Explain the strengths and weaknesses of the sampling method you have chosen.

Data and analysis



Show evidence of taking a sample from the population by:

- displaying the sample data using dot plots
- adding box plots to your dot plots
- providing summary statistics.

Discuss, with supporting evidence, the distribution of the two samples by comparing the two dot plots. Support your discussion by referring to the box plots and relevant summary statistics. Your comparison should include reference to at least three of the following:

- central tendency (median and/or mean)
- symmetry (symmetrical, left-skewed, or right skewed)
- spread (compare the interquartile ranges of the two data sets)
- overlap (do the boxes showing where the middle 50 percent of each group lie overlap)
- any unusual features or outliers.

You may like to use the PEEL framework to craft your responses:

P	**Point**	Make a statement about what you can see in the dot plot or box plot.
E	**Explanation**	Explain why you think this.
E	**Evidence**	Use numbers or facts to back up what you are saying.
L	**Link**	Link your statements to the population, the context, your research or to a prediction you made at the start of your investigation.

Construct informal confidence intervals for the medians of each sample, either using the formula *median* ± 1.5 x $\frac{IQR}{\sqrt{n}}$ or software such as NZGrapher. This YouTube video provides an **explanation of how to construct an informal confidence using NZGrapher**.

Interpret the informal confidence intervals, taking care to state that you are "reasonably confident" that the true population median lies somewhere between the lower and upper limits of the confidence interval.

Use the informal confidence intervals to draw a conclusion about the population parameters of the two groups. Base your conclusion on whether the two confidence intervals overlap (no evidence for a difference between the two medians) or don't overlap (evidence that a difference exists).

Discuss sampling variability, including the variability of estimates. Make the point that another sample could give different results because no two samples are identical.

Assessment tip:

Remember that discussing sampling variation is a requirement of the assessment task. You need to clearly demonstrate that you understand that no two samples from a population are identical.

Conclusion

Write a brief summary of what you discovered in your investigation. Make sure that there is a clear link between your conclusion and the purpose of your investigation, as stated in your introduction.

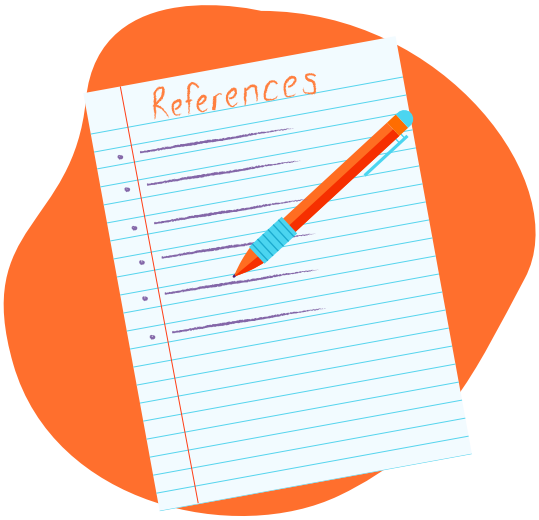
Relate your findings to the background information you found about your topic.

Evaluate the process you went through, discussing possible assumptions, limitations or improvements that could be made. If possible, suggest related areas of interest that could be investigated further.

References

Provide a list of any sources used in your research or referred to in your discussion.

Complete Topic 6 Activity 1
in the Student Practice Booklet



Before moving on to the assessment task, check that you understand:

- [ ] How to structure your report
- [ ] The content you need to include in your report.