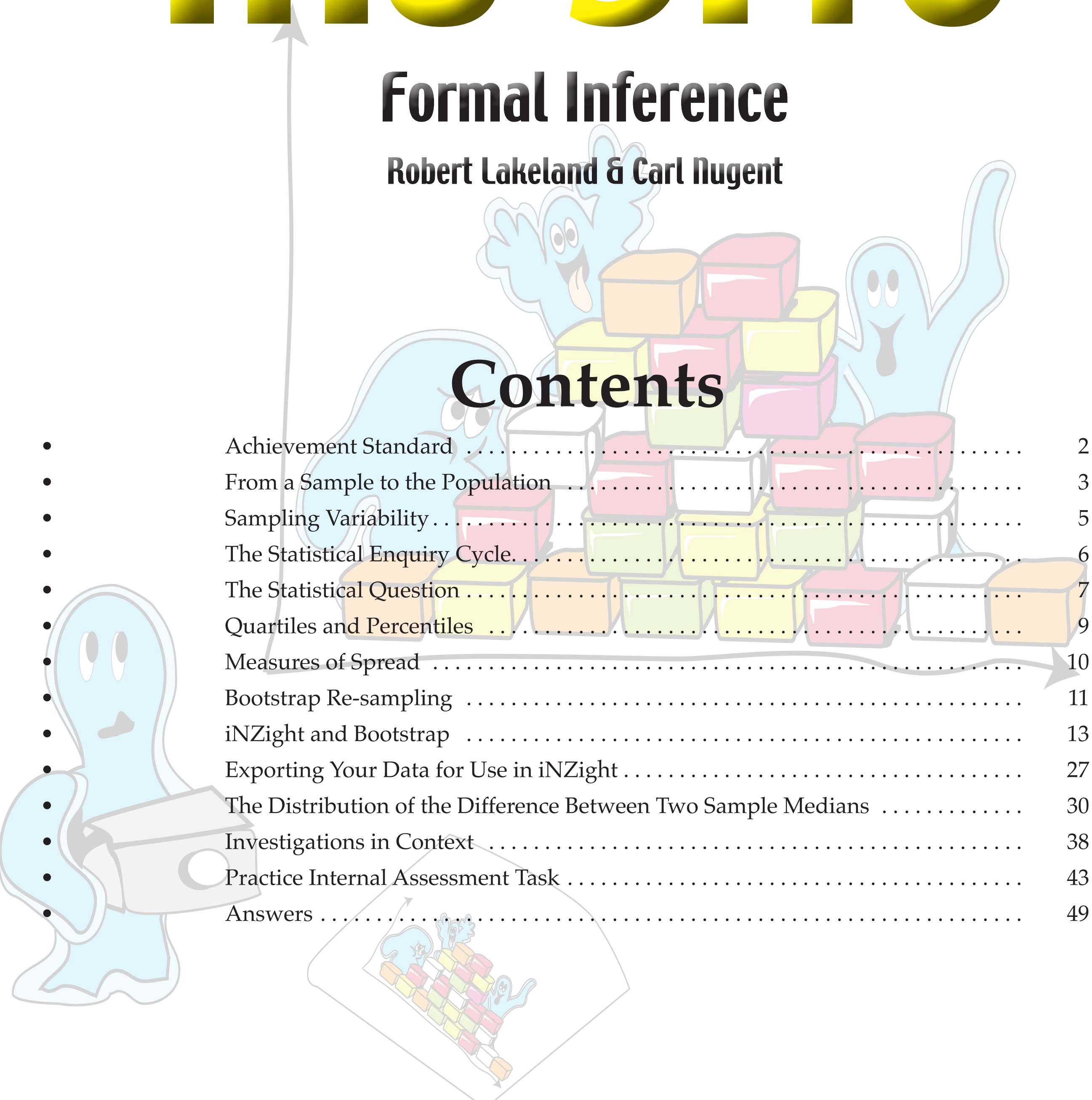
Year 13 Mathematics

IAS 3.10

Formal Inference

Robert Lakeland & Carl Nugent



Contents

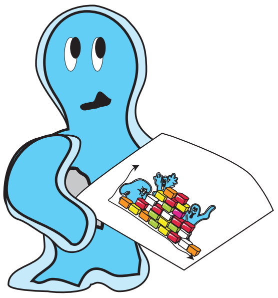
nuLake Ltd
Innovative Publisher of Mathematics Texts

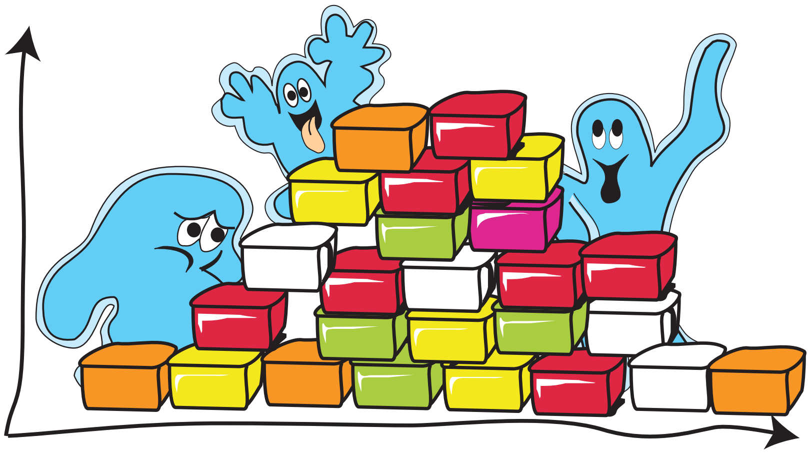
NCEA 3 Internal Achievement Standard 3.10 – Formal Inference

This achievement standard involves using statistical methods to make a formal inference.

Achievement	Achievement with Merit	Achievement with Excellence
• Use statistical methods to make a formal inference.	• Use statistical methods to make a formal inference, with justification.	• Use statistical methods to make a formal inference, with statistical insight.

- This achievement standard is derived from Level 8 of The New Zealand Curriculum and is related to the achievement objectives.
    - Carry out investigations of phenomena, using the statistical enquiry cycle:
        - using existing data sets
        - seeking explanations
        - using informed contextual knowledge, exploratory data analysis, and statistical inference
        - communicating findings and evaluating all stages of the cycle.
    - Make inferences from surveys or experiments:
        - determining estimates and confidence intervals for differences
        - use methods such as re-sampling to assess the strength of the evidence.
- Use statistical methods to make a formal inference involves showing evidence of using each component of the statistical enquiry cycle.
- Use statistical methods to make a formal inference, with justification involves linking components of the statistical enquiry cycle to the context, and/or to the populations, and referring to evidence such as sample statistics, data values, or features of visual displays in support of statements made.
- Use statistical methods to make a formal inference, with statistical insight involves integrating statistical and contextual knowledge throughout the statistical enquiry cycle, and may include reflecting about the process; considering other relevant explanations.
- Using the statistical enquiry cycle to make a formal inference involves:
    - posing a comparison investigative question using a given multivariate data set
    - selecting and using appropriate displays and summary statistics
    - discussing sample distributions
    - discussing sampling variability, including the variability of estimates
    - making an appropriate formal statistical inference
    - communicating findings in a conclusion.





From a Sample to the Population

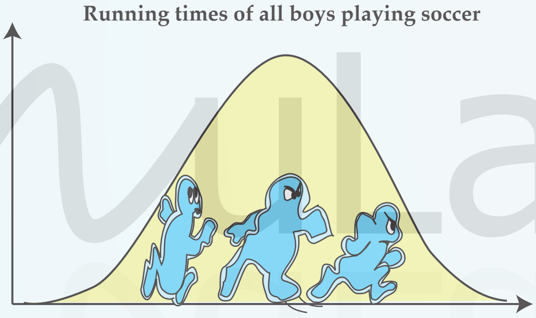
Inference from a Sample

In NCEA 2 you investigated populations by looking at a sample from the population and attempted to infer whether any information or differences you observed in the sample existed in the population.

You may, for example, be looking at the running times of 16 year old male students. You look at a statistic of the running times in your sample (such as the median or mean) and infer that the times of the corresponding parameter in your population is likely to be within a range either side of this sample statistic. We call this range the **confidence interval**.

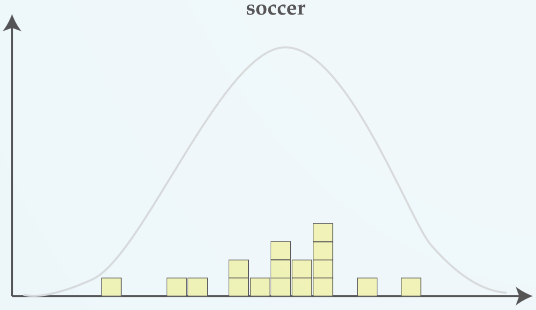
In NCEA 3 you make a formal inference about the difference in a population parameter between two groups back in the population. You use a method known as bootstrap re-sampling. For example, you may wish to examine the median (or mean) running time of boys that play soccer.

Running times of all boys playing soccer

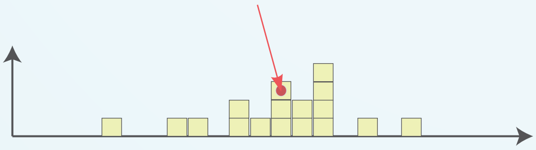


From the population we are working with we select a single random sample.

A single sample of the running times of boys that play soccer



How similar the sample distribution is to the population will depend upon the sample size. From this sample we can calculate sample statistics such as the median.



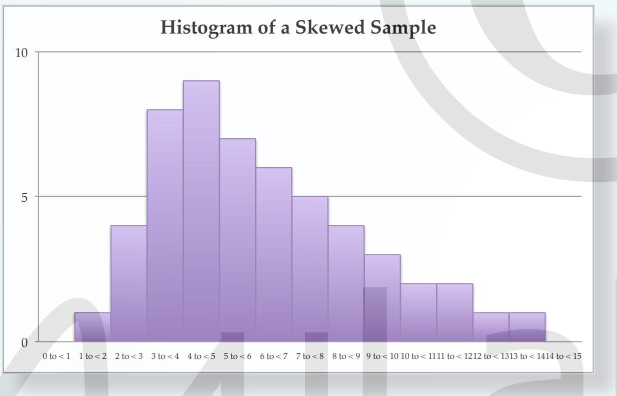
Making an Inference from a Sample to a Population

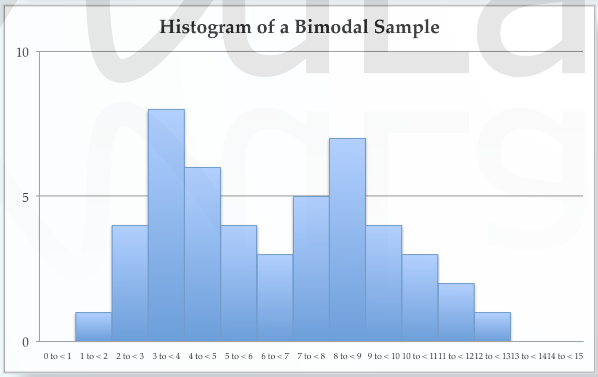
- A random sample is a snapshot of the population. We can use it to find some sample statistics.
- A random sample depends upon chance so it is not a mini copy of the population but, depending upon its size, the sample distribution should be similar to the population distribution and can be used to answer questions about the population.
- We can calculate a confidence interval for a statistical difference between groups and infer that a population parameter difference will very likely occur in that interval.
- The accuracy of our prediction will depend upon
  - the sample size
  - how spread out the population data is
  - the type (shape) of the distribution.

Bootstrap Re-sampling

Modelling the Distribution of Sample Statistics

The statistics we get from our sample (median, mean, quartiles etc.) will be close to the population parameters (median, mean, quartiles etc.). The population parameters are likely to be in a range around the sample statistics called a confidence interval. Statisticians use models to find this interval. Historically statisticians used the Central Limit Theorem model but it is limited to means and requires that either the sample must be large (generally $n \geq 30$) or that the population is approximately normally distributed (i.e. a bell shaped curve symmetrical about the mean). The Central Limit Theorem is not suitable for skewed or bimodal distributions.


Histogram of a Skewed Sample


Histogram of a Bimodal Sample

In 1979 Bootstrap re-sampling was developed and by the 1990s this new technique was preferred as a method for modelling the distribution of sample statistics. Bootstrap re-sampling involves taking a large number of samples (called re-samples) **with** replacement from the original sample and using these large number of re-samples to make inferences about the population that the original sample was taken from.

Bootstrap re-sampling does not make any assumptions about the population distribution and uses this sample distribution as a good estimate of the population distribution. Bootstrap re-sampling can also infer other parameters (median, quartiles etc.) as well as the mean.

Bootstrap re-sampling is the preferred model for making inferences back in the population.

This strange name came from the saying that 'when all else fails you pull yourself up by your own bootstraps'.

Obviously it is impossible to pull one's self up by your own bootstraps but the phrase has come to mean to succeed without any outside help.

Applying this to sampling, we do not know the original population distribution, but work entirely with the sample to make our inferences about the population.

Bootstrap re-sampling can be used to generate inferences for any parameter. It is not limited to only means, you can make inferences for medians, quartiles etc.

Example

A quality control officer was checking the weights of bags of kiwifruit. They wanted to know the 95% percentile confidence interval for the population mean and median. The sample in grams is

1 – 5	6 – 10	11 – 15	16 – 20
416	407	404	426
410	401	368	417
405	415	415	421
385	387	398	423
415	406	416	392

This file is available in the 'IAS 3.10 Data Files' folder which can be downloaded from the NuLake website under 'Downloads', 'Year 13', 'IAS 3.10'. The file is called Kiwifruit.csv.

The company advertises the bags as 400 g but plans to pack them to an average weight of 410 g.

Use Bootstrap re-sampling to generate the 95% percentile confidence interval and then with justification comment on whether the sample confirms the company's expectations.

Explain why bootstrap re-sampling is the best approach for this sample.

Using the **iNZight module.**

The sample has the following sample statistics:

Mean = 406.4 g   Median = 408.5 g

Using 1000 Bootstrap re-samples, with the **iNZight module**, the sample statistics are

Mean = 406.4 g   Median = 409.5 g

From iNZight the 95% percentile bootstrap confidence interval for the population mean is

399.6 g to 412.2 g

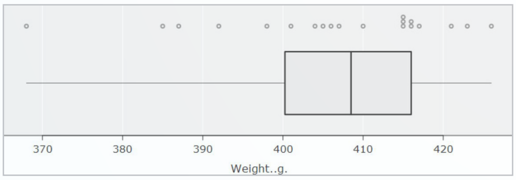
From iNZight the 95% percentile bootstrap confidence interval for the population median is

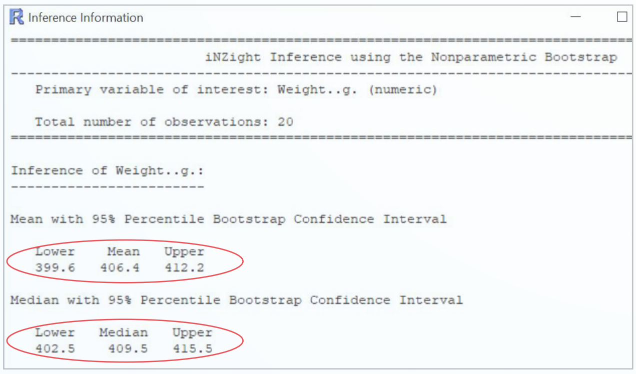
402.5 g to 415.5 g

The company's average packing weight is within our confidence interval so may be correct.

The company advertises the bags as 400 g but it is possible the population mean is as low as 399.6 g. As this figure rounds to 400 g the company label is also justified.

The small sample size and skewed distribution means that bootstrap re-sampling is appropriate.





```
==========================================================================
                   iNZight Inference using the Nonparametric Bootstrap
--------------------------------------------------------------------------
   Primary variable of interest: Weight..g. (numeric)

   Total number of observations: 20
==========================================================================

Inference of Weight..g.:
------------------------

Mean with 95% Percentile Bootstrap Confidence Interval

   Lower    Mean    Upper
   399.6   406.4    412.2

Median with 95% Percentile Bootstrap Confidence Interval

   Lower   Median    Upper
   402.5    409.5    415.5
```

Using the **iNZight VIT - Bootstrap confidence interval construction module.**

The sample has the following sample statistics:

Mean = 406.4 g   Median = 408.5 g

Using 1000 Bootstrap re-samples, with the **VIT Bootstrap confidence interval construction module**, the sample statistics are

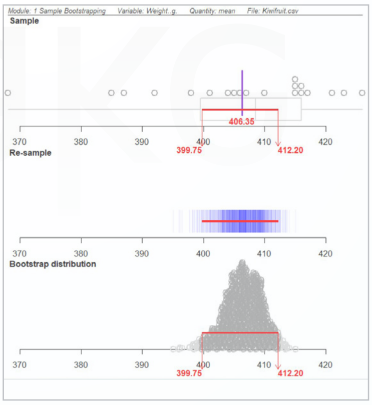
Mean = 406.4 g   Median = 408.5 g

From iNZight the 95% percentile bootstrap confidence interval for the population mean is

399.8 g to 412.2 g

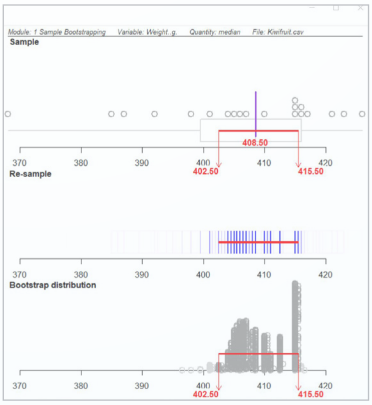
From iNZight the 95% percentile bootstrap confidence interval for the population median is

402.5 g to 415.5 g



The company advertises the bags as 400 g but it is possible the population mean is as low as 399.8 g. As this figure rounds to 400 g the company label is also justified.

The small sample size and skewed distribution means that bootstrap re-sampling is appropriate.



5. The fuel economy of a particular model of car is to be estimated. A sample of 20 cars, of the designated model, is taken and their petrol consumption recorded (litres per 100 km). The results were

9.2, 8.6, 7.9, 10.4, 6.6, 11.3, 7.8, 9.5, 9.1, 8.9, 10.2, 7.7, 6.9, 8.2, 9.7, 10.0, 8.1, 7.9, 9.3, 7.7.

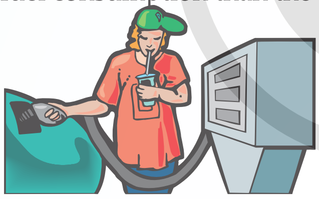
The file for this question is called Petrol.csv.

a) Use bootstrap re-sampling to find a 95% percentile confidence interval for the mean petrol consumption of this model of car.

b) Use bootstrap re-sampling to find a 95% percentile confidence interval for the median petrol consumption of this model of car.

c) Sketch, on the grid, the distribution of means with the confidence interval shown.

d) Explain why in this case the mean is a better measure of the fuel consumption than the median.



6. The waiting time at a particular doctor's surgery is to be estimated. The waiting time of a sample of 30 of the doctor's patients is recorded (in minutes). The results were:

8.7, 5.9, 10.6, 4.3, 13.5, 4.1, 5.9, 8.6, 26.5, 2.6, 2.1, 18.3, 7.5, 11.3, 6.1, 9.2, 7.4, 3.6, 31.2, 8.6, 4.7, 6.3, 4.1, 5.4, 9.7, 7.9, 21.6, 14.3, 13.5, 1.1.

The file for this question is called Waiting.csv.

a) Use bootstrap re-sampling to find a 95% percentile confidence interval for the mean waiting time at the doctor's surgery.

b) Use bootstrap re-sampling to find a 95% percentile confidence interval for the median waiting time at the doctor's surgery.

c) Is the mean or median waiting time best used as a measure of the waiting time in this case?

d) Sketch, on the grid, the distribution of medians with the confidence interval shown.

e) The surgery wants to advise patients of the typical waiting times to see the doctor. What time(s) should they use?

The Distribution of the Difference Between Two Sample Medians

If we have two samples that we wish to compare we could plot the box and whisker plot of both samples on the same graph and see if the difference between the medians is likely to be reflected in the population.

In the example below of a sample of 71 girls and 95 boys the boys' median hand size is 171.4 mm and girls' median hand size is 169.8 mm. Can we conclude from our sample that back in the population the boys' median hand size is greater than the girls' median hand size?

Our 95% percentile confidence interval for the boys' median hand size is from 169 to 175 mm and for the girls' median is from 169 to 171 mm. Note that each time you complete bootstrap re-sampling you will get a slightly different interval as it is based on chance.

These two confidence intervals overlap. Therefore it seems unlikely that back in the population the boys' median would be bigger than the girls' median but we need to investigate this.

What we need to look at is the distribution of **differences** of the sample medians.

With the iNZight software we can take bootstrap samples from the population of boys' and girls' hand sizes and plot a distribution of sample differences between the boys' median hand size and the sample of girls' median hand sizes.

The data file for this example is available in the 'IAS 3.10 Data Files' folder which can be downloaded from the NuLake website under 'Downloads', 'Year 13', 'IAS 3.10'. The data file for this example is called Hand Size.csv.

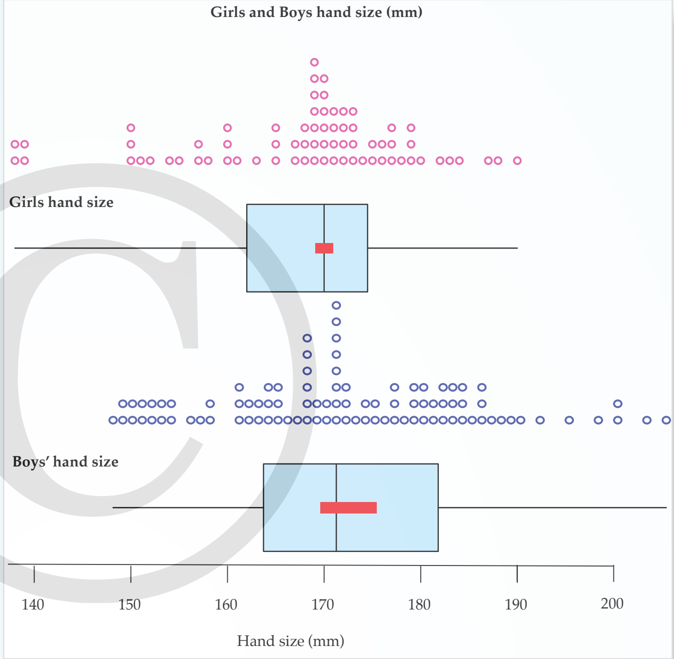
Select the Bootstrap confidence interval construction module in the Visual Inference Tools (VIT) and import the data file 'Hand Size.csv'.

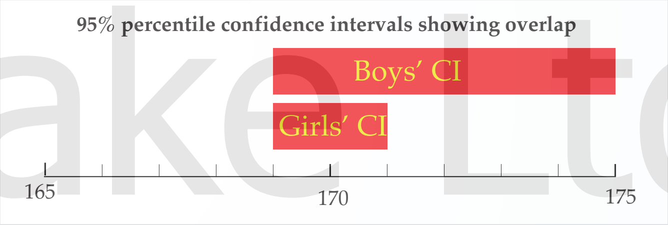
Select 'Gender' as Variable 1 and 'Hand size' as Variable 2.

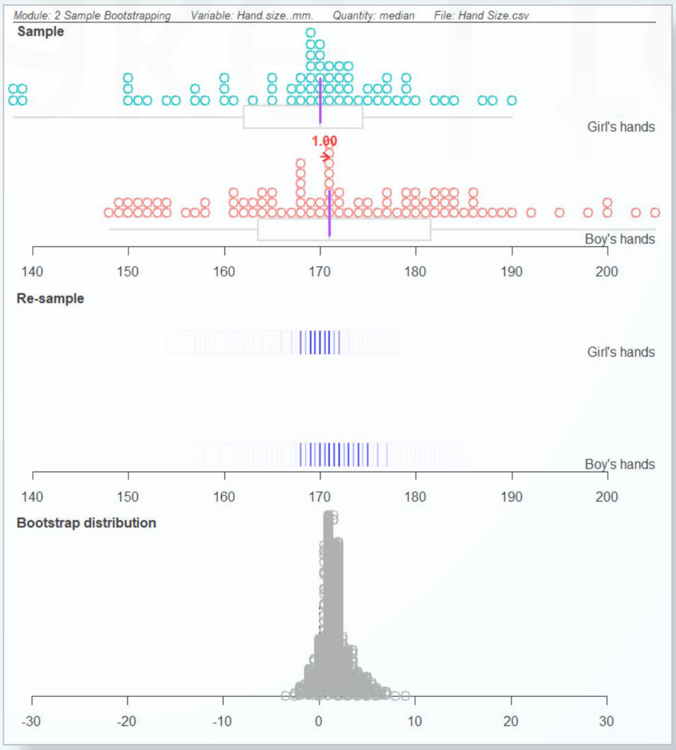
Go to Analyse and at the top of the screen is the population distribution for hand sizes of each sex. Select **median** and 'Record my choices' and run 1000 samples for 'Include bootstrap distribution'.

The difference between the sample medians is plotted as a dot plot on the right.

We can see that sometimes the two sample medians differed by as much as 10 mm and often the difference was negative (i.e. the girls' median size was bigger than the boys' median).






Module: 2 Sample Bootstrapping Variable: Hand.size..mm. Quantity: median File: Hand Size.csv

Answers

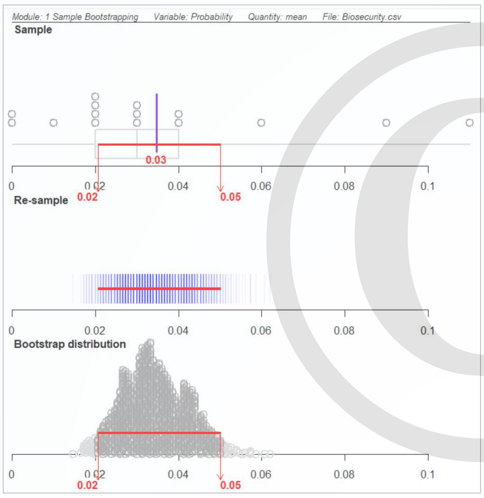
Results in most questions will vary slightly as they rely on bootstrap re-sampling.

Page 22

1. a) 95% percentile confidence interval for the mean by bootstrap re-sampling is 0.021 to 0.052.

b) No, the confidence interval goes beyond 0.05 so they are not justified in saying they are less than 0.05.

c)



d) The sample is only 15 elements and is not a normal distribution

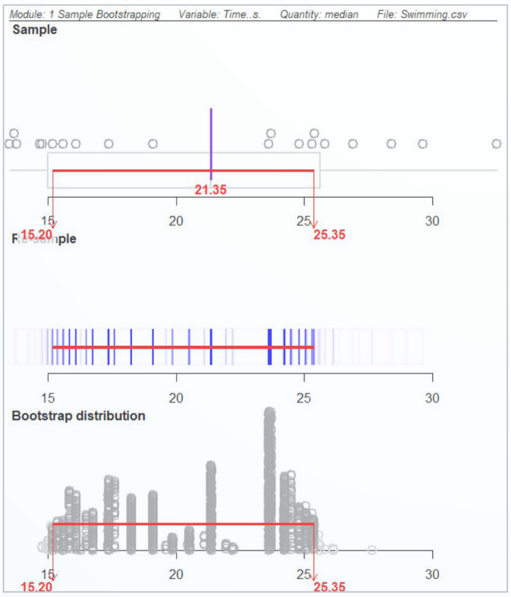
2. a) Unusual distribution, small sample size and information required about the median.

b) 95% percentile confidence interval for the mean by bootstrap re-sampling is 18.5 seconds to 23.6 seconds (1 dp).

95% percentile confidence interval for the median by bootstrap re-sampling is 15.2 seconds to 25.4 seconds (1 dp).

c) The bimodal distribution means the median swings greatly depending on which end of the distribution is over represented in the bootstrap sample.

d)



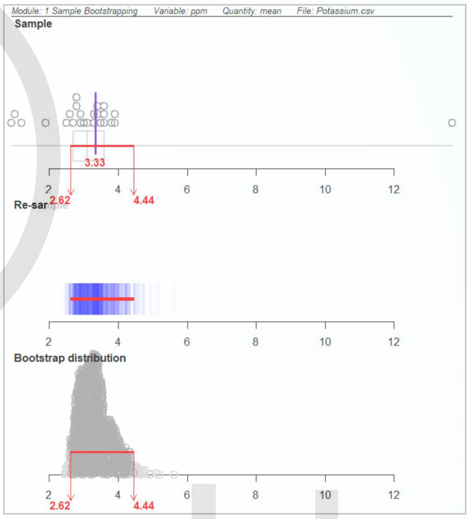
Page 23

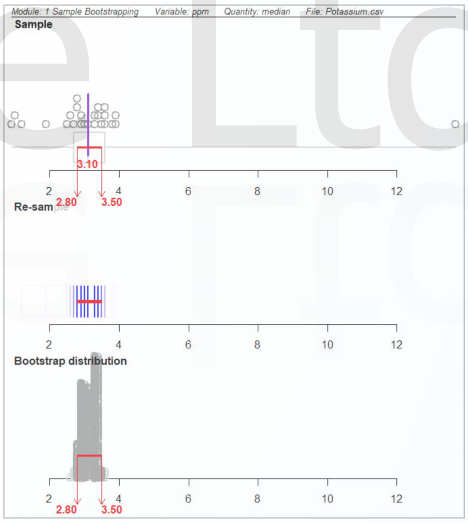
3. a) 95% percentile confidence intervals by bootstrap re-sampling are:
Mean 2.62 ppm to 4.44 ppm
Median 2.80 to 3.50 ppm.

b) The extreme value of 13.7 ppm affects the mean resulting in high means every time it is selected and even higher if it is selected more than once.

c) The extreme value of 13.7 ppm does not affect the median as the median is the middle number.

d) and e)





4. a) The 95% percentile confidence interval for the mean by bootstrap re-sampling is 1.49 to 2.21 mg/l.
The 95% percentile confidence interval for the median by bootstrap re-sampling is 1.60 to 2.40 mg/l.

b) Both the mean and median of the population could be over 2 mg/l. The school is not compliant at the 95% percentile level with the Ministry of Education requirement.

c) The distribution is not normal with observations spread erratically between 0.5 and 2.8 mg/l. The population distribution is likely to be similar to the sample distribution.