Statistics 3.10

Relating to page 109 of *Level 3 Statistics Learning Workbook*

Sampling variability and sample size

Every time we take a sample from a population, we will get slightly different results – in other words, there will always be **sampling variability** – so that the sample distribution will differ from the population distribution.

There are two reasons that a statistic calculated using a sample (such as the sample median) differs from the actual population parameter:

- **Sampling error** – every sample will contain a sampling error simply because statistics are worked out using data from a sample, rather than by using the whole population. So each sample median is only an estimate of the actual population median, and so is very likely to be different from the true population median.
- **Non-sampling error** – non-sampling errors may also be present due to the sampling methods used, which may introduce bias into the sample.

The effect of sampling variability can be observed using iNZight.

In the following example, repeated samples of size 30 are taken from a population of data, and the differences between sample medians are recorded for each re-sample.

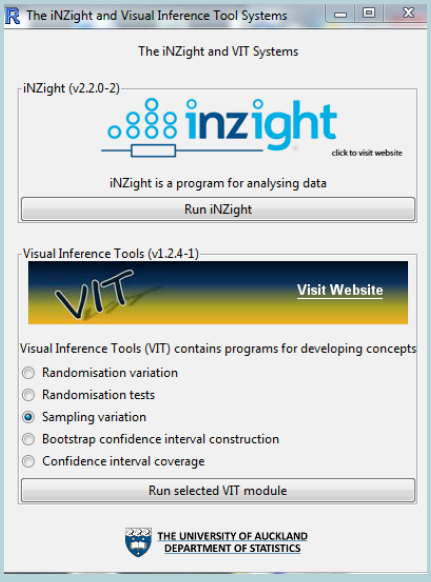
Example

The dataset 'Salt in soups' is treated as the population.

This data file is available at ESA RESOURCES

Using iNZight, select **Sampling variation**

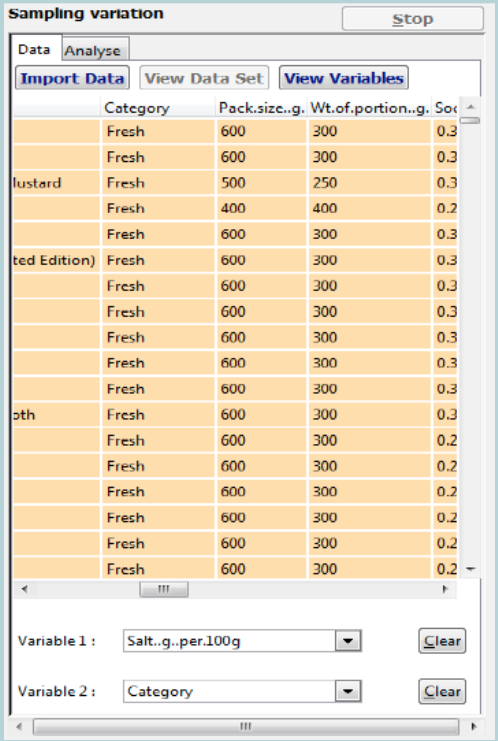
Run selected VIT module



Import the file **Salt in soups**

Select the variable 'Salt..g..per..100.g' as Variable 1

Select the variable 'Category' as Variable 2



Now take repeated samples of size 30 from the data set:

Select **Analyse**

Quantity **Median**

Sample size **30**

Record my choices

In the 'population' of all soups in this dataset, the difference between the median salt contents is 0.10 g/100 g, with ready-to-serve soups having the higher median salt level.

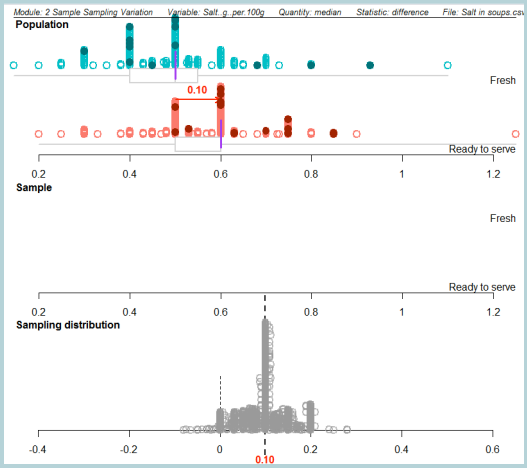
Sampling variability will now be observed by taking repeated samples of size 30, and recording the difference between the medians for each sample. To do this, select:

Include sampling distribution

Number of repetitions 1 000

Go

The sampling distribution shows the values of the difference in sample medians for each resample.



Most samples have differences between median salt contents of around 0.10 g per 100 g, with the differences in sample median salt contents having values from approximately –0.09 g/100 g (where the median salt level is lower in ready-to-eat soups) to +0.28 g/100 g.

The variation between samples (sampling error) is due to calculating medians using a sample of size 30 each time, rather than calculating the medians using the entire population (for which there is a fixed difference between medians of 0.10 g/100 g).

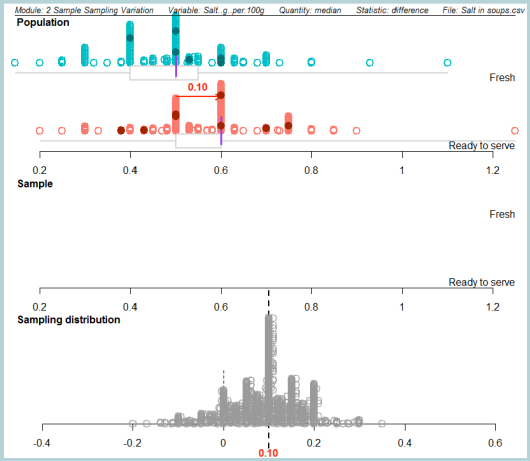
Sample size affects sampling variability:

- As sample size decreases, sampling variability increases and there will be an increase in the range of values in the sampling distribution (i.e. there will be a wider interval of differences in sample medians).
- As sample size increases, sampling variability decreases and there will be a decrease in the range of values in the sampling distribution (i.e. there will be a narrower interval of differences in sample medians).

The effect of sample size on sampling variability will be illustrated by repeating the process in the example above, but using sample sizes of 15 and 100.

Example

1. Taking repeated samples of size 15 from the data set gives the following distribution of differences between sample medians.

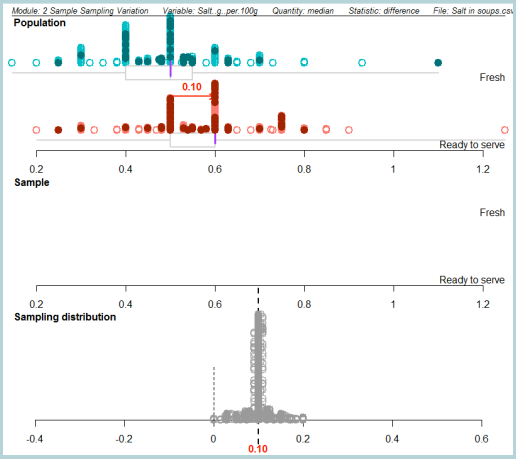


Sample size 15

Once again, most samples have differences between median salt contents of around 0.10 g per 100 g, with the differences in sample median salt contents having values from approximately –0.2 g/100 g to +0.35 g/100 g (compared with a narrower range of –0.09 g/100 g to 0.28 g/100 g for sample size 30).

2. Taking repeated samples of size 100 from the data set gives the following distribution of differences between sample medians.

Sample size 100



For samples of size 100, differences in sample median salt contents for fresh and ready-to-serve soups are centred around 0.10 g/100g, and take on values between 0 g/100 g and 0.2 g/100 g (a range of only 0.2 g, compared with a range of 0.37 g for samples of size 30 and a range of 0.55 g for samples of size 15).

Comparing the sampling distributions, it can be seen that the differences between sample medians is less variable for larger samples.

As would be expected, the distribution of larger samples will more closely approximate the population distribution, while a small sample risks being unusual just by chance, and is more likely to be a sample that is not representative of the population.

In your report you MUST include a reflection on sampling variability.

- Each time a sample is taken, different sample statistics could be found.
- As a result a slightly different confidence interval could be generated.

You should reflect on whether it is likely that taking another sample and hence obtaining a different

confidence interval would change your conclusion. This could include discussion on the distance of the limits of the confidence interval from zero (for example, if the limits of the original 95-percentile confidence interval were well away from zero, it is unlikely that another sample would produce a confidence interval that did contain zero and hence alter the conclusion).

Exercise F: Sampling variability and sample size

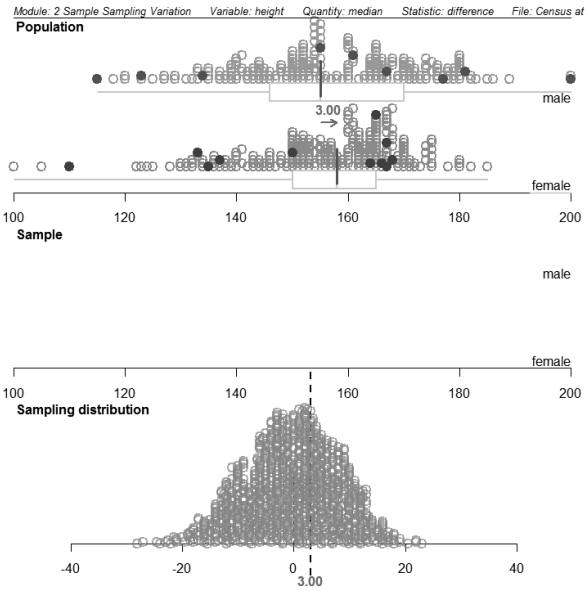
1. Import the data file Census at School-500.csv, from the data folder of iNZightVIT. Treating this data file as the population, use the Sampling variation option of iNZightVIT to investigate the effect of sample size on sampling variability. Do this by taking 1 000 re-samples of the following sizes from the data, and comparing the differences in sample median heights for boys and girls for each re-sample.
   a. Do this for the following sample sizes:
      i. sample size 20
      ii. sample size 50
      iii. sample size 100
   b. Compare and comment on your results.

2. Repeat the exercise in question 1 using a different variable over two groups in the data set.

Answers

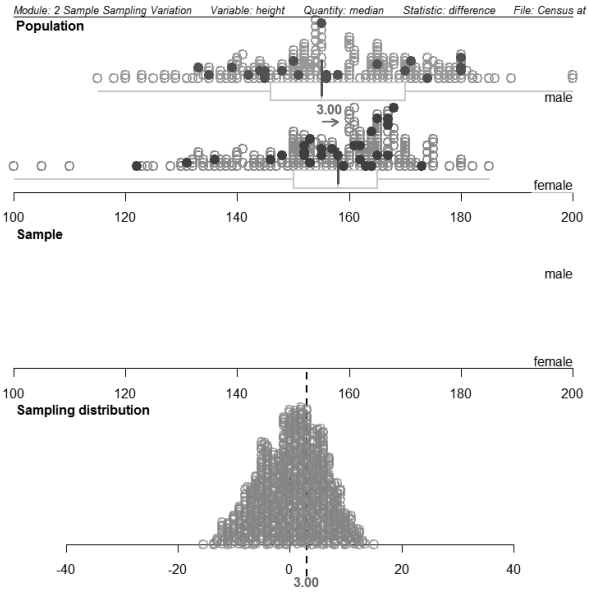
1. *Answers will vary – an example is given.*

   a. i. The following graph shows the distribution of the differences in median heights for males and females for 1 000 re-samples of size 20.



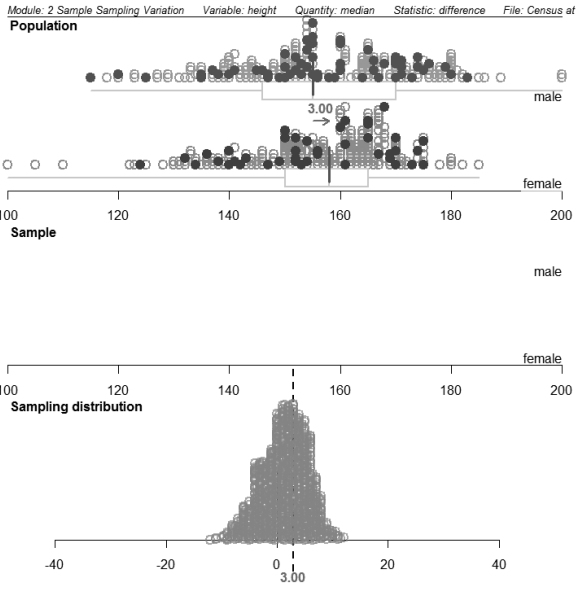
   For samples of size 20, differences between male and female median heights lay between –30 cm and 25 cm approximately.

   ii. The following graph shows the distribution of the differences in median heights for 1 000 re-samples of size 50.



   For samples of size 50, the differences between male and female median heights lay between –18 cm and 17 cm approximately.

   iii. The following graph shows the distribution of the differences in median heights for 1 000 re-samples of size 100.



   For samples of size 100, the differences between male and female median heights lay between –12 cm and 10 cm approximately.

   b. For this population, the difference between the population median heights of males and females is 3 cm. All three distributions of sample differences in median heights were centred around this value. However, the width of the distribution varied, depending on the size of the samples.

   - For re-samples of size 20, the differences between male and female median heights had a range of about 55 cm.
   - For re-samples of size 50, the differences differences between male and female median heights had a range of about 35 cm.
   - For samples of size 100, the differences between male and female median heights had a range of about 22 cm.

   Clearly, larger sample sizes reduce sampling variability.

2. *Answers will vary.*