LAB 02

ENVX2001 Applied Statistical Methods

Table of contents

Exercise 1 – Soil carbon (walk-through) 2
  Calculating the 95% confidence interval of the mean . 2
  Doing the calculations in R 3
Exercise 2 - Soil carbon: what if…? 5
Exercise 3 - Kangaroos 6
  Should we return to the same locations? 6
  Different locations 7
  Same locations 8
  Putting it together 9
Exercise 4 – Equivalence to t-tests 9
Thanks! 10
  Attribution 10

> **i** Learning outcomes
>
> At the end of this practical students should be able to:
>
> - explain conceptually the benefits of re-sampling (or not) the same units for monitoring studies.
> - use R to create rudimentary sampling designs;
> - estimate means and associated CIs for simple random designs;
> - estimate means and associated CIs for stratified random designs;
> - use R for estimating the change in means (and associated CIs) for monitoring schemes when the sample units are (i) resampled (ii) not resampled.

Exercise 1 – Soil carbon (walk-through)

Karunaratne et al. (2012)[1] measured soil carbon (%) in the Cox's creek catchment in northern NSW using a stratified random sampling scheme. The results are summarised below.

- **USE** is the land use type;
- **N** is the number of samples collected;
- **MEAN** is the mean soil carbon content;
- **VARIANCE** is the variance of the mean;
- **PERCENT_AREA** is the percentage of the catchment area represented by each land use type.

Recall that the confidence interval for the mean is an effective way to communicate the precision of the estimates (how well we know the true mean), and is derived from the standard error of the mean.

[1]Karunaratne S. B., Bishop T. F. A., Odeh I. O. A., Baldock J. A., Marchant B. P. (2014) Estimating change in soil organic carbon using legacy data as the baseline: issues, approaches and lessons to learn. *Soil Research* **52**, 349-365.

Calculating the 95% confidence interval of the mean

The *weighted* mean (y_s) and variance ($Var(\bar{y}_s)$) of the mean for a stratified random sample are given by:

Weighted mean:

1. Calculate mean value per stratum
2. Multiply each mean value by the weight of the stratum
3. Sum the weighted mean values to get the overall mean

$$\bar{y}_s = \sum_{i=1}^{L} \bar{y}_i \times w_i$$

where L is the number of strata, $\bar{y}_i$ is the mean of stratum i, and w_i is the weight of stratum i, and

$$Var(\bar{y}_s) = \sum_{i=1}^{L} w_i^2 \times Var(\bar{y}_i)$$

where $Var(\bar{y}_i)$ is the variance of the mean for stratum i.

We can then calculate the 95% CI for the mean:

$$95\%CI = \bar{y}_s \pm t_{n-L}^{0.025} \times \sqrt{Var(\bar{y}_s)}$$

where L = number of strata, n = total number of samples, $t_{n-L}^{0.025}$ is the t-critical value for the 95% CI and $\sqrt{Var(\bar{y}_s)}$ is the variance of the mean for the stratified random sample.

Is stratified random sampling a good idea for this study? Let's find out.

Weighted variance of the mean:

1. Calculate the variance of the mean for each stratum
2. Multiply the variance of the mean by the **square** of the weight of the stratum
3. Sum all the weighted variances to get the overall variance of the mean

95% CI for the mean:

- Weighted mean **plus** (t-critical value × standard error of the mean)
- Weighted mean **minus** (t-critical value × standard error of the mean)

> **!** Important
>
> **Understanding how to calculate the 95% CI using summary statistics is examinable**. If you find it difficult to understand the mathematical notation, it might be easier to rewrite them in your own words – for example, see the side notes. We will never ask you to calculate values directly, but will ask you to interpret the results.

Doing the calculations in R

The data is provided below. The first is the `stratified` dataset, which contains the mean and variance of the soil carbon content for each land use type.

```r
stratified <- data.frame(
    use = c(
        "Forest", "Dryland Cropping", "Pasture-Vertosol",
         ↪  "Pasture-Other",
        "Irrigated"
    ),
    n = c(9, 14, 14, 2, 5),
    mean = c(0.65, 0.96, 1.06, 1.31, 0.78),
    variance = c(1.4, 0.4, 0.8, 0.4, 0.8),
    percent_area = c(20, 35, 35, 6, 4)
)
```

You should copy and paste this data into your own document and run `stratified` to view the data.

Look at the code and see if you can understand how a data frame is created in R. The `data.frame` function is used to create a data frame, and the `c` function is used to create vectors of data. Each vector forms a column in the data frame.

Question 1 Estimate the mean and associated 95% CI assuming **stratified** random sampling using the `carbon` dataset. We can use the `weighted.mean` function to calculate the weighted mean, and the `qt` function to calculate the t-critical value for a 95% CI. You should find out what these functions do by seeking their help documentation using the `?` operator.

```r
?weighted.mean
?qt
```

The standard error of the mean can be calculated using the following:

```r
# SE: square_root(sum(weighted variance of the mean))
```

Finally, the 95% CI can be calculated using the following:

```r
# 95% CI = weighted mean ± (t_crit * SE)
```

Tip: you can manipulate the object and subset specific rows, columns or cells using `$` and `[]` respectively. For example, to access the `mean` column of the `carbon` dataset, you can use `carbon$mean`. To access the mean value for the `Forest` stratum, you can use `carbon$mean[1]`.

Exercise 2 - Soil carbon: what if...?

What if the authors had used simple random sampling? Below is the data object `overall` which contains the overall mean and variance of the soil carbon content for the entire catchment. It is also the last row of the table above.

```
overall <- data.frame(
    use = "Overall",
    n = 44,
    mean = 0.92,
    variance = 1,
    percent_area = 100
)
```

Calculating the 95% CI for the mean using simple random sampling is similar to the stratified random sampling. The only difference is that the variance of the mean is calculated using the overall variance and sample size.

$$Var(\bar{y}) = \frac{Var(y)}{n}$$

where $Var(y)$ is the variance of the entire dataset and n is the total number of samples.

Thus the 95% CI for the mean is:

$$95\%CI = \bar{y} \pm t^{0.025}_{n-1} \times \sqrt{Var(\bar{y})}$$

where n is the total number of samples, and $t^{0.025}_{n-1}$ is the t-critical value for the 95% CI.

Question 2 Estimate the mean and associated 95% CI assuming **simple random sampling** using the `overall` dataset, recalling that:

```
# 95% CI = mean ± (t_crit * SE)
```

THE UNIVERSITY OF SYDNEY

Question 3 Would you recommend stratified random sampling for future surveys? Provide evidence from the results of questions 2 and 3.

Exercise 3 - Kangaroos

Should we return to the same locations?

In this exercise we will explore the impact that resampling the locations has on the precision (width of the 95% CI) with which we estimate the change in mean between 2 surveys. We will also show the equivalence of this to 2-sample t-tests.

The key equation for estimating the variance of the change in mean is:

$$Var(\Delta\bar{y}) = Var(\Delta\bar{y}_2) + Var(\Delta\bar{y}_1) - 2 \times Cov(y_1, y_2)$$

where:

- $Var(\Delta\bar{y})$ is the variance of the change in mean between 2 surveys;
- $Var(\Delta\bar{y}_2)$ is the variance of the mean for the 1st survey (baseline);
- $Var(\Delta\bar{y}_1)$ is the variance of the mean for the 2nd survey (repeat);
- $Cov(y_1, y_2)$ is the covariance between the means of the 2 surveys.

When we resample the same locations, we include the covariance term which describes the relationship between the observations in the 2 surveys. If resample at **different locations** we assume the **covariance is equal to 0**, that is:

$$Var(\Delta\bar{y}) = Var(\Delta\bar{y}_2) + Var(\Delta\bar{y}_1)$$

In this exercise we will consider a study where kangaroos (number/km^2) were counted in a woodland in an initial survey, and then 2 years later. The data is below.

```
'data.frame':   5 obs. of  2 variables:
 $ baseline: num  4 2 6 1 3
 $ rerun   : num  12 11 13 14 9
```

Can you generate a `data.frame` object from the output above? Try it yourself first, and then click on show the code below to see the answer (Note: PDF output will not hide the answer).

```
baseline <- c(4, 2, 6, 1, 3)
rerun <- c(12, 11, 13, 14, 9)
kangaroos <- data.frame(baseline, rerun)
```

Question 4 Intuitively, what would be the impact on the precison with which we estimate the mean when:

(a) we resample the same locations in the repeat survey?
(b) we sample at different locations in the repeat survey?

Different locations

Let's assume that the resampled locations are completely different from the baseline survey. This means that the covariance term is equal to 0.

Question 5 Use R to estimate the 95% CI for the change in mean between the 2 surveys. The equation is:

$$95\%\ CI = \Delta\left(\bar{y}\right) \pm t_{df}^{0.025} \times \sqrt{Var\left(\Delta\bar{y}\right)}$$

You will need to calculate $Var\left(\Delta\left(\bar{y}_1\right)\right)$, $Var\left(\Delta\left(\bar{y}_2\right)\right)$, and $Cov\left(\bar{y}_1, \bar{y}_2\right)$.

As a starting point R script for calculating the covariance and variance for the **observations** is below, as is the code to calculate the t critical value ($t_{df}^{0.025}$). When we sample at different locations between surveys:

$$df = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$$

```r
# If you want to isolate data from the two sampling periods:
t0 <- kangaroos$baseline
t2 <- kangaroos$rerun

# NOTE: functions below are not assigned to objects.
# To calculate covariance:
cov(t0, t2)

# To calculate variances:
var(t0)
var(t2)

# t-critical value for a 95% CI with 8 degrees of freedom:
qt(0.975, 8)
```

Same locations

Now let's assume that the resampled locations are the same as the baseline survey. This means that the covariance term is *not* equal to 0.

Question 6 Assuming we resampled the **same locations** as the baseline survey, use R to estimate the 95% CI for the change in mean between the 2 surveys.

Hint: the following formulae are used:

$$\Delta\bar{y} = \bar{y}_2 - \bar{y}_1$$

$$Var(\Delta\bar{y}) = Var(\bar{y}_2) + Var(\bar{y}_1) - 2 \times Cov(\bar{y}_1, \bar{y}_2)$$

Note: when we sample at same locations between surveys the df = number of paired sites - 1 = 5 - 1 = 4.

Question 7 Was there a significant change in the mean number of kangaroos between the study period, if we assume the same locations were resampled, or if different locations were sampled?

Putting it together

As you should realise by now, the same data *may* produce different results depending on how it was analysed. This is the **danger** of statistics: **not clearly understanding the assumptions and implications of an experimental design can lead to incorrect analysis and conclusions.**

Any statistical software, including R, will not tell you if you have made a mistake in your analysis. It is up to you to understand the implications of your design and to ensure that your work is appropriate.

In th case of this kangaroo study, the results are the same for both scenarios, which means that the implications lie in the context of future studies.

Question 8 In the future, would you re-sample the same sites, or visit new sites in a kangaroo monitoring program of the same woodland?

Exercise 4 – Equivalence to t-tests

The scenario when we *resample at different locations* between surveys is actually the equivalent to a **two-sample t-test**, and the scenario when we *resample the same locations* is equivalent to a **paired t-test**.

Without manually calculating and comparing the 95% CI, we can use the `t.test()` function in R to perform both of these tests.

Check your answers in R by trying to perform both of these using the `t.test()` function. Assume the variances are equal for the two-sample t-test (use the argument `var.equal = TRUE`for this).

Question 9 Compare the 95% CI for each scenario between the `t.test()` results and the ones you calculated manually. Are they same?

Thanks!

Did you know you can also knit to PDF? Check the documentation for R Markdown or Quarto for more information.

Attribution