Lab 5 - Experimental Design

ENVX2001 Applied Statistical Methods

Semester 1, 2026

> 💡 Learning outcomes
>
> At the end of this lab students should be able to:
>
> - distinguish between sampling units and experimental units;
>
> - use R to generate randomisations for CRDs and RCBDs;
>
> - use R to analyse experiments with blocking, and assess the usefulness of blocking.
>
> All of the data for this lab is in the **Data5.xlsx** file.

Exercise 1 - Randomisation for a CRD using R (Walk-through)

Consider a glass house experiment conducted on a bench on which it was judged that the growing conditions would be consistent across the bench. The experimenter had five different smoked water treatments for assisting the germination of *Banksia* seeds. She had 50 dishes on which she could place seeds randomly selected from a uniform batch of *Banksia* seeds. The dishes were to be placed on the bench and each of the five water solutions allocated randomly to 10 of the dishes.

Question 1.1

(i) What would the degrees of freedom be in a corresponding ANOVA table?

Solution

- Treatment df = t − 1 = 5 − 1 = 4

- Residual df = N − t = 50 − 5 = 45

- Total df = N − 1 = 50 − 1 = 49

Question 1.2

(ii) Suppose that Treatment 1 was a control. In situations where the comparisons of interest are of the other treatments (2 to 5) with the control, and not amongst the other treatments, it can be advantageous to increase the number of control replicates to 20. What are the degrees of freedom in the corresponding ANOVA table?

Solution

- Treatment df = t − 1 = 5 − 1 = 4
- Residual df = N − t = 60 − 5 = 55
- Total df = N − 1 = 60 − 1 = 59

Question 1.3

(iii) The formula for standard error of the difference (SED) is:

$$SED = \sqrt{Resid\ MS\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Given the SED formula, what percentage improvement (i.e. reduction) has been achieved for the SED between Treatment 1 and 2 when there are equal replicates *(i)* versus unequal replicates *(ii)*? Assume the Residual MS is a constant.

Solution

For the equal replicate design:

$$SED = \sqrt{Resid\ MS\left(\frac{1}{10} + \frac{1}{10}\right)} \approx 0.447 \times \sqrt{Resid\ MS}.$$

For the second design:

$$SED = \sqrt{Resid\ MS\left(\frac{1}{20} + \frac{1}{10}\right)} \approx 0.387 \times \sqrt{Resid\ MS}.$$

So SED(extra control rep) / SED(equal rep) = 0.387 / 0.447 = 0.866, i.e. a 13.4% reduction in the SED.

Exercise 2 - The balance between the number of sampling units and experimental units (Walk-through)

Horticulturalists have been studying the nutrient requirements of lettuce growing in a sand medium; in particular they are looking at the nitrate concentration in leaf petioles in response to varying applied nitrogen nutrient levels (5, 11, 18 and 32 mmol/L). Lettuce plants are grown in separate pots, so different nitrogen levels may be applied to individual pots, and one or more leaves sampled per plant. They are also interested in optimising their experimental protocol for future studies; consequently they conducted five separate experiments where they sampled different number of leaves per plant, and differing number of plants, but kept the total number of leaves sampled at 128 in each experiment. The experiments were conducted as follows:

- Experiment 1: 16 leaves per plant; 2 plants per treatment;
- Experiment 2: 8 leaves per plant; 4 plants per treatment;
- Experiment 3: 4 leaves per plant; 8 plants per treatment;
- Experiment 4: 2 leaves per plant; 16 plants per treatment;
- Experiment 5: 1 leaf per plant; 32 plants per treatment.

Similar lettuce plants were used across all five experiments, and in all cases, a CRD was used to allocate the plants amongst the treatments. The datasets are located in the *Data5.xlsx* file. There are separate sheets for each Experiment called *Experiment1*, *Experiment2*, etc.

Question 2.1

(i) What are the experimental units and what are the sampling units in the above experiments. For each experiment, how many experimental units and how many sampling units are used?

Solution

The experimental unit is the plant, and the sampling unit is the leaf on the plant. All experiments have 128 sampling units (leaves). However, the number of experimental units varies, and is: Experiment 1: 8; Experiment 2: 16; Experiment 3: 32; Experiment 4: 64; and Experiment 5: 128.

Question 2.2

(ii) When there are multiple sampling units per experimental unit, in simple situations (as is here), an appropriate method is simply to average the responses over all the sampling units belonging to an experimental unit, and analyse these. So in this case, we need to average the

observed nitrate values over all the leaves in the plant. The code below shows how to do this using a function called `rowMeans` which finds the average in a row across multiple columns. Here we do this for *Experiment1*.

```
CODE
#
```

Solution

```
CODE
library(readxl)
exp1←read_excel("data/Data5.xlsx",sheet="Experiment1")
exp1$treatment←as.factor(exp1$treatment)
exp1$response←rowMeans(exp1[,3:18])
str(exp1)
```

```
OUTPUT
tibble [8 × 19] (S3: tbl_df/tbl/data.frame)
 $ treatment: Factor w/ 4 levels "N1","N2","N3",..: 1 1 2 2 3 3 4 4
 $ plant    : num [1:8] 1 2 3 4 5 6 7 8
 $ leaf1    : num [1:8] 2.46 1.73 3.31 2.49 2.4 2.6 3.06 2.23
 $ leaf2    : num [1:8] 2.45 1.3 2.5 2.73 2.56 2.42 1.95 2.39
 $ leaf3    : num [1:8] 3.24 2.37 2.02 2.32 1.67 2.48 2.38 2.69
 $ leaf4    : num [1:8] 2.36 1.46 1.97 1.54 1.78 2.3 1.93 2.4
 $ leaf5    : num [1:8] 3.17 1.52 2.46 2.5 2 2.09 1.32 1.74
 $ leaf6    : num [1:8] 2.87 1.99 1.52 2.62 3.32 1.66 1.87 2.84
 $ leaf7    : num [1:8] 2.39 2.79 2.07 2.71 1.75 2.46 2.42 2.36
 $ leaf8    : num [1:8] 2.94 2.03 1.68 1.92 2.49 2.68 2.22 3.3
 $ leaf9    : num [1:8] 2.76 2.23 2.68 3.07 2.21 2.07 2.38 1.62
 $ leaf10   : num [1:8] 3 2.17 1.45 2.48 2.33 1.75 2.76 1.82
 $ leaf11   : num [1:8] 2.54 2.74 1.96 2.15 1.46 2.61 2.04 3.03
 $ leaf12   : num [1:8] 3.19 1.36 2.26 2.73 2.2 2.63 2.07 3.66
 $ leaf13   : num [1:8] 3.03 1.45 1.94 2.22 1.61 2.08 2.09 4.03
 $ leaf14   : num [1:8] 2.42 1.19 2.38 2.37 2.6 3.44 2.18 1.59
 $ leaf15   : num [1:8] 2.68 1.12 2.66 1.75 1.3 2.46 2.18 2.81
 $ leaf16   : num [1:8] 2.56 0.77 1.85 1.49 1.61 2.5 1.95 2.27
 $ response : num [1:8] 2.75 1.76 2.17 2.32 2.08 ...
```

Question 2.3

(iii) Perform a one-way ANOVA on these average data. (It will also be helpful to obtain the treatment means using the `emmeans` function, as done in previous topics. What are your conclusions about the effect of varying applied nitrogen levels? You don't have to look at post-hoc tests, just examine the F-test.

```
CODE
#
```

Solution

```
CODE
mod1←aov(response~treatment,data=exp1)
summary(mod1)
```

```
OUTPUT
            Df Sum Sq Mean Sq F value Pr(>F)
treatment    3 0.0208 0.00693   0.045  0.986
Residuals    4 0.6186 0.15466
```

```
CODE
library(emmeans)
emmeans(mod1,"treatment")
```

```
OUTPUT
 treatment emmean    SE df lower.CL upper.CL
 N1          2.26 0.278  4     1.49     3.03
 N2          2.24 0.278  4     1.47     3.02
 N3          2.23 0.278  4     1.46     3.01
 N4          2.36 0.278  4     1.59     3.13

Confidence level used: 0.95
```

Conclusion: There are no significant differences in mean nitrate level across the four nitrogen treatment groups (F = 0.045; df = 3,4; P = 0.99).

Question 2.4

(iV) Repeat this for Experiments 2 to 5. What are your conclusions in each experiment?

Experiment 2: 8 leaves per plant, 4 plants per treatment

```
CODE
#
```

Solution

```
CODE
exp2←read_excel("data/Data5.xlsx",sheet="Experiment2")
exp2$treatment←as.factor(exp2$treatment)
exp2$response←rowMeans(exp2[,3:10])
mod2←aov(response~treatment,data=exp2)
summary(mod2)
```

```
OUTPUT
            Df Sum Sq Mean Sq F value Pr(>F)
treatment    3 0.3042  0.1014   0.911  0.465
Residuals   12 1.3360  0.1113
```

```
CODE
library(emmeans)
emmeans(mod2,"treatment")
```

```
OUTPUT
 treatment emmean    SE df lower.CL upper.CL
 N1          1.96 0.167 12     1.60     2.33
 N2          2.13 0.167 12     1.76     2.49
 N3          2.35 0.167 12     1.99     2.71
 N4          2.11 0.167 12     1.75     2.47

Confidence level used: 0.95
```

Experiment 3: 4 leaves per plant, 8 plants per treatment

```
CODE
#
```

Solution

```
CODE
exp3←read_excel("data/Data5.xlsx",sheet="Experiment3")
exp3$treatment←as.factor(exp3$treatment)
exp3$response←rowMeans(exp3[,3:6])
mod3←aov(response~treatment,data=exp3)
summary(mod3)
```

```
OUTPUT
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    3  1.889  0.6296   5.825 0.00318 **
Residuals   28  3.026  0.1081
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
CODE
library(emmeans)
emmeans(mod3,"treatment")
```

```
OUTPUT
 treatment emmean    SE df lower.CL upper.CL
 N1          1.92 0.116 28     1.68     2.16
 N2          2.23 0.116 28     1.99     2.47
 N3          2.44 0.116 28     2.21     2.68
 N4          2.56 0.116 28     2.32     2.80

Confidence level used: 0.95
```

Experiment 4: 2 leaves per plant, 16 plants per treatment

```
CODE
```

```
#
```

Solution

```
CODE
exp4←read_excel("data/Data5.xlsx",sheet="Experiment4")
exp4$treatment←as.factor(exp4$treatment)
exp4$response←rowMeans(exp4[,3:4])
mod4←aov(response~treatment,data=exp4)
summary(mod4)
```

```
OUTPUT
           Df Sum Sq Mean Sq F value  Pr(>F)
treatment   3  4.136  1.3787   5.761 0.00157 **
Residuals  60 14.357  0.2393
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
CODE
library(emmeans)
emmeans(mod4,"treatment")
```

```
OUTPUT
 treatment emmean    SE df lower.CL upper.CL
 N1          2.03 0.122 60     1.79     2.27
 N2          2.11 0.122 60     1.87     2.36
 N3          2.15 0.122 60     1.91     2.40
 N4          2.68 0.122 60     2.43     2.92

Confidence level used: 0.95
```

Experiment 5: 1 leaf per plant, 32 plants per treatment. Note, no averaging required as sampling unit = experimental unit.

```
CODE
#
```

Solution

```
CODE
exp5←read_excel("data/Data5.xlsx",sheet="Experiment5")
exp5$treatment←as.factor(exp5$treatment)
exp5$response←exp5$leaf1
mod5←aov(response~treatment,data=exp5)
summary(mod5)
```

```
OUTPUT
```

```
           Df Sum Sq Mean Sq F value Pr(>F)
treatment    3   2.97  0.9915   3.219 0.0251 *
Residuals  124  38.19  0.3080
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CODE
```
library(emmeans)
emmeans(mod5,"treatment")
```

OUTPUT
```
 treatment emmean     SE  df lower.CL upper.CL
 N1           2.00 0.0981 124     1.81     2.20
 N2           2.14 0.0981 124     1.94     2.33
 N3           2.31 0.0981 124     2.12     2.51
 N4           2.40 0.0981 124     2.20     2.59

Confidence level used: 0.95
```

While Experiment 2 also did not indicate significant differences in mean nitrate level across the four nitrogen treatments, significant differences were detected in the subsequent experiments, particularly in Experiments 3 and 4. There is a trend of increasing mean nitrate concentrations with increasing nitrogen treatment, indicating a dose-response relationship.

Question 2.5

(v) Summarise the results in terms of the balance between sampling plants versus leaves, and what recommendations would you make for future studies?

CODE
```
#
```

Solution

With a fixed number of leaves (sampling units), increasing the number of plants and reducing the number of leaves sampled per plant results in an increased significance for the treatment effect being detected. However, there is evidence that only sampling one leaf per plant results in a moderate decline in the significance levels. These findings are seen when either the Residual MS, F or P-values are used to make this assessment. So for this scheme, it would be recommended to use Experiment 3 or Experiment 4 protocols. Note that the standard errors of means (SE) also decline with increasing plants per treatments, though this changes little with eight or more.

Experiment	Plants_Per_Treatment	Leaves_Per_Plant	Resid_MS	F_Stat	P_Value	SE
1	2	16	0.155	0.04	0.986	0.278
2	4	8	0.111	0.91	0.465	0.167
3	8	4	0.108	5.82	0.003	0.116
4	16	2	0.239	5.76	0.002	0.122
5	32	1	0.308	3.22	0.025	0.098

Question 2.6

(vi) There are two sources of random variation encountered in these experiments:

-variation between leaves within a plant; and

-variation between plants within the same treatment group.

How do you think your recommendations might change in *(iv)* if:

-there was (virtually) no variation between leaves within a plant; only between plants; or

-there was (virtually) no variation between plants, only variation between leaves within a plant.

Solution

If there is virtually no variation between leaves within a plant, then just using one leaf per plant would be sufficient, so Experiment 5 would be the optimal protocol here.

If there is much leaf to leaf variation, but little between-plant variation, then it is best to increase the number of leaves sampled per plant, and use fewer plants. So Experiment 1 might be optimal here (but this would give the very minimal level of replication, so for a cautionary approach, Experiment 2 might be used).

Exercise 3 - A paired t-test generalises to a one-way ANOVA – RCBD

Fifteen farms cooperated in a field trial in which a normal fattening ration for pigs (ration A) was compared with the same rations supplemented with a small trace of copper (ration B). Each farmer set up two pens of pigs, as similar as possible in all respects, and allocated the two rations

at random, ration A to one pen and ration B to the other. The mean weight gains per pen (g/day) are stored in the two columns of the **copper** sheet in the **Data5.xlsx** file.

This description clearly suggests the design is an RCBD, with 15 farms but only 2 treatments. As we have seen, a paired design is the simplest form of RCBD. In this example we will not worry about model assumptions.

Question 3.1

(i) Firstly analyse the data using a paired t-test. Assuming you read the data is as a data frame called `copper` the code is `t.test(x=copper$RationA,y=copper$RationB,paired=T)`. Interpret the results.

```
CODE
#
```

Solution

```
CODE
library(readxl)
copper←read_excel("data/Data5.xlsx",sheet = "copper")
head(copper)
```

```
OUTPUT
# A tibble: 6 × 3
   Farm RationA RationB
  <dbl>  <dbl>   <dbl>
1    1     422     531
2    2     526     467
3    3     476     558
4    4     499     585
5    5     422     472
6    6     503     522
```

```
CODE
t.test(x=copper$RationA,y=copper$RationB,paired=T)
```

```
OUTPUT

    Paired t-test

data:  copper$RationA and copper$RationB
t = -3.2545, df = 14, p-value = 0.005761
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -67.90951 -13.95716
sample estimates:
mean difference
      -40.93333
```

Pigs fed Ration B (copper supplement) have a significantly higher weight gain than those on Ration A (P = 0.006). On average, the cupper supplemented pigs (B) have a weight gain of

41 g/day more than the control, but a 95% confidence interval for this difference is between 14.0 and 67.9 g/day more.

Question 3.2

(ii) Next we will obtain an analysis of the data using ANOVA. To do this analysis however, you will first need to re-arrange the data set into a form suitable for ANOVA. That is, you will need a single column of weight gains, a column of treatment identifiers, and a column indicating the Farm number. The code below will do this for you:

- `dat←stack(copper[,2:3])`

- `farms←rep(c(1:15),times=2)`

- `coppern←cbind(farms,dat)`

The first line stacks (using the `stack` function) the 2 columns of weight gain on top of each other (column name is `values`) and adds a column indicating which ration is associated with each weight gain value (column name is `ind`).

The second line repeats (using the `rep` function) the values of farm numbers twice so for each of the weight gains we have an associated farm.

The third line joins the two data frames together (using the `cbind` function) to create the new stacked dataset which can be used by the `aov` function.

```
CODE
head(copper)
```

```
OUTPUT
# A tibble: 6 × 3
   Farm RationA RationB
  <dbl>   <dbl>   <dbl>
1     1     422     531
2     2     526     467
3     3     476     558
4     4     499     585
5     5     422     472
6     6     503     522
```

```
CODE
dat←stack(copper[,2:3])
farms←rep(c(1:15),times=2)
coppern←cbind(farms,dat)
str(coppern)
```

```
OUTPUT
'data.frame':   30 obs. of  3 variables:
 $ farms : int  1 2 3 4 5 6 7 8 9 10 ...
 $ values: num  422 526 476 499 422 503 445 449 299 517 ...
 $ ind   : Factor w/ 2 levels "RationA","RationB": 1 1 1 1 1 1 1 1 1 1 ...
```

Question 3.3

(iii) The general code for performing an 1-way anova with a randomised complete block design is:

```
aov(response~block+treatment,data=data)
```

In this example the `response` is the weight gain, `blocks` are farms and the `treatment` is ration. Your names will be different. Having obtained the analysis, check that the results from the ANOVA is identical to the paired t-test:

- the same *P*-value;
- the same test statistic (though expressed as an *F*-statistic instead of a *t*-statistic, with $F = t^2$ to that from the paired *t*-test.

```
CODE
#
```

Solution

```
CODE
block.anova←aov(values~as.factor(farms)+ind,data=coppern)
anova(block.anova)
```

```
OUTPUT
Analysis of Variance Table

Response: values
                 Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(farms) 14  71524  5108.9   4.306 0.004963 **
ind               1  12567 12566.5  10.592 0.005761 **
Residuals        14  16610  1186.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparisons with paired t-test.

- P = 0.006 in both analyses
- $t^2 = (-3.25)^2 = 10.59 = F$
- Residual df = 14 = df for t-test

Exercise 4 - One-way ANOVA - RCBD

Three diets for hamsters were tested for differences in weight gain after a specified period of time. Six inbred lines were used with three hamsters selected from each line. The three diets were assigned at random to the hamsters in each line. The data is in the **hamsters** sheet in the **Data5.xlsx** file.

Question 4.1

(i) Are the mean weight gains the same across diets? Are any diets more or less effective than other diets? You will need to use the emmeans package and its emmeans() function. Read in the data and perform post-hoc tests.

```
CODE
#
```

Solution

```
CODE
library(readxl)
hamster←read_excel("data/Data5.xlsx",sheet="hamster")
head(hamster)
```

```
OUTPUT
# A tibble: 6 × 3
  Inbred_Line Wt_Gain  Diet
  <chr>         <dbl> <dbl>
1 A                18     1
2 B                16     1
3 C                17     1
4 D                19     1
5 E                12     1
6 F                13     1
```

```
CODE
hamster$Diet←as.factor(hamster$Diet)
hamster$Inbred_Line←as.factor(hamster$Inbred_Line)
block.anova←aov(Wt_Gain~Inbred_Line+Diet,data=hamster)
anova(block.anova)
```

```
OUTPUT
Analysis of Variance Table

Response: Wt_Gain
            Df Sum Sq Mean Sq F value   Pr(>F)
Inbred_Line  5 71.167  14.233  7.4912 0.003650 **
Diet         2 36.333  18.167  9.5614 0.004773 **
Residuals   10 19.000   1.900
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Perform post-hoc tests.

```
CODE
library(emmeans)
emmeans(block.anova, pairwise ~ Diet)
```

```
OUTPUT
$emmeans
 Diet emmean    SE df lower.CL upper.CL
 1      15.8 0.563 10    14.6    17.1
 2      13.0 0.563 10    11.7    14.3
 3      12.7 0.563 10    11.4    13.9

Results are averaged over the levels of: Inbred_Line
Confidence level used: 0.95


$contrasts
 contrast       estimate    SE df t.ratio p.value
 Diet1 - Diet2    2.833 0.796 10   3.560  0.0131
 Diet1 - Diet3    3.167 0.796 10   3.979  0.0067
 Diet2 - Diet3    0.333 0.796 10   0.419  0.9087

Results are averaged over the levels of: Inbred_Line
P value adjustment: tukey method for comparing a family of 3 estimates
```

There are significant differences in mean weigh gain of hamsters amongst the three diets (F = 9.56; df = 2, 10; P = 0.005). Based on the post-hoc tests, hamsters fed diet 1 had a significantly higher weight gain than those fed diets 2 or 3, whereas there was no significant difference between weights for hamsters on Diets 2 and 3.

Question 4.2

(ii) Re-run the analysis without blocking, i.e. as a CRD. Are the conclusions different? Would you use blocking in the future? What proportion of the variation was explained by blocking?

```
CODE
#
```

Solution

```
CODE
crd.anova←aov(Wt_Gain~Diet,data=hamster)
anova(crd.anova)
```

```
OUTPUT
Analysis of Variance Table
```

```
Response: Wt_Gain
          Df Sum Sq Mean Sq F value  Pr(>F)
Diet       2 36.333 18.1667  3.0222 0.07892 .
Residuals 15 90.167  6.0111
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now have a non-significant F-test. Based on Block SS in the original analysis value of 71.167, we can see that the majority of the Residual SS (90.167) in the CRD is from the between-block differences. By including the block effect through the RCBD we have dramatically reduced the Residual SS.

The proportion of variation explained by the Blocking is:

$\frac{Block\ SS}{Total\ SS} = \frac{71.167}{126.5} = 0.56$.

Therefore 56% of the variation in the response is explained by the blocking. This is compelling evidence for using blocking in future versions of the experiment.