

ENTMLGY 6707 Entomological Techniques and Data Analysis

Supplemental activity (do not submit): Randomized Complete Block and Latin Squares

1 Introduction

Stratified designs are used to (ideally) increase precision of estimates by experimentally controlling or reducing variation. This tutorial covers how to analyze two types of experimental designs: randomized complete block and Latin squares. The examples include multiple ways of analyzing the same exact data.

Nowadays, some folks use mixed-effects models to analyze stratified designs. Be aware that a fixed-effect only approach might be preferred, given that we assume random effects are normally distributed and it is hard to test that assumption when there are only a few levels of a random effect (e.g., 3 blocks).

Either way, you still have to be careful about specifying the random effects correctly. And you will notice that if you do, the sums of squares and F -statistics for each treatment are typically equivalent. There are sometimes differences, but rarely do they influence the overall conclusions.

The following packages are necessary to complete this tutorial.

```
library(car)
library(lme4)
library(lmerTest)
library(tidyverse)
library(agricolae)
library(emmeans)
```

2 Randomized complete block

The data “Trefoil” contains data from seven genetically different populations of birdsfoot trefoil (forage crop) seedlings evaluated for their response to a single application of a herbicide. The experimental unit was a plot containing 6 plants (= sample units) of a chosen population, and there were 8 replicates (blocks) in an RCB layout. The data collected were individual plant fresh weights (in grams) three weeks after the herbicide treatment.

```
trefoil <- read.table("Trefoil.txt", header=T, sep="\t",
                     colClasses = c("factor","factor", "numeric",
                                     "numeric", "numeric", "numeric",
                                     "numeric", "numeric", "numeric"))
head(trefoil)
```

	Rep	Sample	Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7
1	1	1	0.060	0.238	0.296	0.246	0.318	0.550	0.321
2	1	2	0.243	0.215	0.141	0.484	0.322	0.474	0.516
3	1	3	0.142	0.107	0.346	0.359	0.341	0.521	0.640
4	1	4	0.213	0.109	0.613	0.173	0.351	0.525	0.559
5	1	5	0.055	0.251	0.208	0.144	0.168	0.580	0.364
6	1	6	0.038	0.322	0.354	0.141	0.369	0.400	0.508

```
summary(trefoil)
```

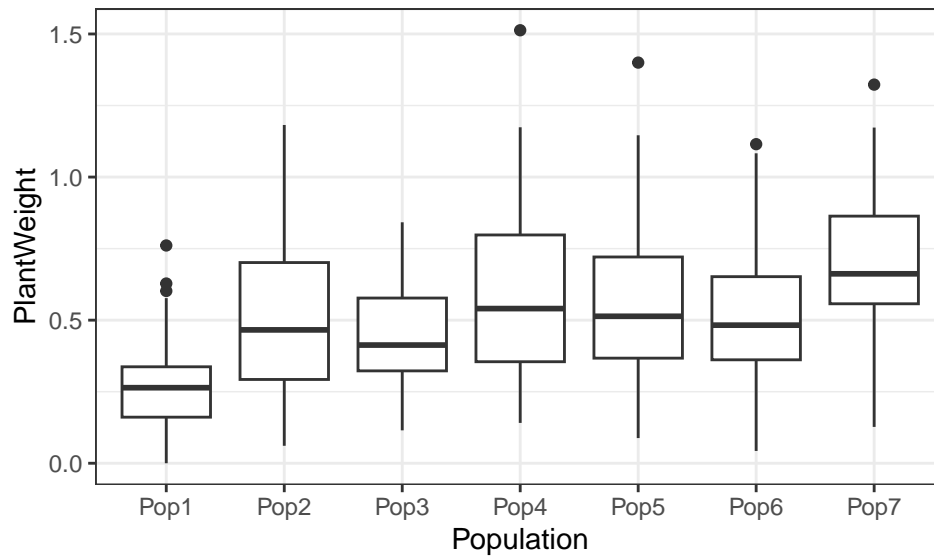
	Rep	Sample	Pop1	Pop2	Pop3
1	: 6	1:8	Min. :0.0000	Min. :0.0610	Min. :0.1150
2	: 6	2:8	1st Qu.:0.1610	1st Qu.:0.2928	1st Qu.:0.3227
3	: 6	3:8	Median :0.2640	Median :0.4660	Median :0.4130
4	: 6	4:8	Mean :0.2794	Mean :0.4904	Mean :0.4522
5	: 6	5:8	3rd Qu.:0.3370	3rd Qu.:0.7013	3rd Qu.:0.5773
6	: 6	6:8	Max. :0.7610	Max. :1.1820	Max. :0.8420
(Other):12					
	Pop4	Pop5	Pop6	Pop7	
Min.	:0.1410	Min. :0.0880	Min. :0.0430	Min. :0.1270	
1st Qu.	:0.3548	1st Qu.:0.3673	1st Qu.:0.3615	1st Qu.:0.5573	
Median	:0.5405	Median :0.5135	Median :0.4825	Median :0.6620	
Mean	:0.6061	Mean :0.5734	Mean :0.5234	Mean :0.6974	
3rd Qu.	:0.7977	3rd Qu.:0.7208	3rd Qu.:0.6520	3rd Qu.:0.8635	
Max.	:1.5130	Max. :1.4000	Max. :1.1150	Max. :1.3230	

2.1 Wide vs. long format

The data above are in “wide” format: there are several observations on a single line (one each for populations 1-7). R requires “long” format for fitting linear models, and the **tidyverse** has a nice function (**pivot_longer**) that enables us to reformat the data. The below code is creating a new data frame by taking all columns that “starts_with” **Pop** and creating new columns called **Population** and **PlantWeight**, into which the column header text (e.g., **Pop1**) and associated value (e.g., 0.060) are input. For example, the first and second row of the new data will have **Population** values equal to **Pop1** and **Pop2** and **PlantWeight** values of 0.060 and 0.238. For this to work, the column headers for whichever variable you are shifting from wide to long format need to start with a unique string of letters (**Pop** in this case).

```
trefoil_long <- trefoil %>%  
  pivot_longer(  
    cols = starts_with("Pop"),  
    names_to = "Population",  
    values_to = "PlantWeight")  
  
head(trefoil_long)
```

```
# A tibble: 6 x 4  
  Rep Sample Population PlantWeight  
  <fct> <fct>   <chr>         <dbl>  
1 1      1      Pop1           0.06  
2 1      1      Pop2          0.238  
3 1      1      Pop3          0.296  
4 1      1      Pop4          0.246  
5 1      1      Pop5          0.318  
6 1      1      Pop6          0.55
```



2.2 aov()

```
fit_aov_RCB <- aov(PlantWeight ~ Rep + Population, data=trefoil_long)
summary(fit_aov_RCB)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Rep	7	4.432	0.6331	12.54	3.16e-14 ***
Population	6	5.044	0.8407	16.66	< 2e-16 ***
Residuals	322	16.251	0.0505		

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

2.3 lm()

```
fit_lm_RCB <- lm(PlantWeight ~ Rep + Population, data=trefoil_long)
Anova(fit_lm_RCB, type="III")
```

Anova Table (Type III tests)

Response: PlantWeight

	Sum Sq	Df	F value	Pr(>F)
--	--------	----	---------	--------

```

(Intercept)  0.1504    1  2.9808    0.08522 .
Rep          4.4315    7 12.5436 3.155e-14 ***
Population   5.0441    6 16.6572 < 2.2e-16 ***
Residuals    16.2514 322
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

2.4 mixed-effects model

```

fit_lmer_RCB <- lmer(PlantWeight ~ Population + (1|Rep), data=trefoil_long)
anova(fit_lmer_RCB, type=3)

```

```

Type III Analysis of Variance Table with Satterthwaite's method
      Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
Population 5.0441 0.84069      6    322 16.657 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

2.5 pairwise comparisons

```

emmeans(fit_lmer_RCB, pairwise~"Population")

```

```

$emmeans
  Population emmean      SE    df lower.CL upper.CL
Pop1         0.279 0.0528 15.2    0.167    0.392
Pop2         0.490 0.0528 15.2    0.378    0.603
Pop3         0.452 0.0528 15.2    0.340    0.565
Pop4         0.606 0.0528 15.2    0.494    0.718
Pop5         0.573 0.0528 15.2    0.461    0.686
Pop6         0.523 0.0528 15.2    0.411    0.636
Pop7         0.697 0.0528 15.2    0.585    0.810

```

```

Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95

```

```

$contrasts
  contrast      estimate      SE    df t.ratio p.value
Pop1 - Pop2 -0.2111 0.0459 322    -4.603 0.0001

```

Pop1 - Pop3	-0.1728	0.0459	322	-3.768	0.0037
Pop1 - Pop4	-0.3267	0.0459	322	-7.124	<.0001
Pop1 - Pop5	-0.2940	0.0459	322	-6.412	<.0001
Pop1 - Pop6	-0.2440	0.0459	322	-5.320	<.0001
Pop1 - Pop7	-0.4181	0.0459	322	-9.117	<.0001
Pop2 - Pop3	0.0383	0.0459	322	0.835	0.9812
Pop2 - Pop4	-0.1156	0.0459	322	-2.521	0.1551
Pop2 - Pop5	-0.0830	0.0459	322	-1.809	0.5426
Pop2 - Pop6	-0.0329	0.0459	322	-0.718	0.9915
Pop2 - Pop7	-0.2070	0.0459	322	-4.514	0.0002
Pop3 - Pop4	-0.1539	0.0459	322	-3.356	0.0153
Pop3 - Pop5	-0.1212	0.0459	322	-2.644	0.1165
Pop3 - Pop6	-0.0712	0.0459	322	-1.552	0.7127
Pop3 - Pop7	-0.2453	0.0459	322	-5.349	<.0001
Pop4 - Pop5	0.0326	0.0459	322	0.712	0.9918
Pop4 - Pop6	0.0827	0.0459	322	1.804	0.5466
Pop4 - Pop7	-0.0914	0.0459	322	-1.993	0.4214
Pop5 - Pop6	0.0501	0.0459	322	1.092	0.9303
Pop5 - Pop7	-0.1240	0.0459	322	-2.704	0.1004
Pop6 - Pop7	-0.1741	0.0459	322	-3.796	0.0033

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 7 estimates

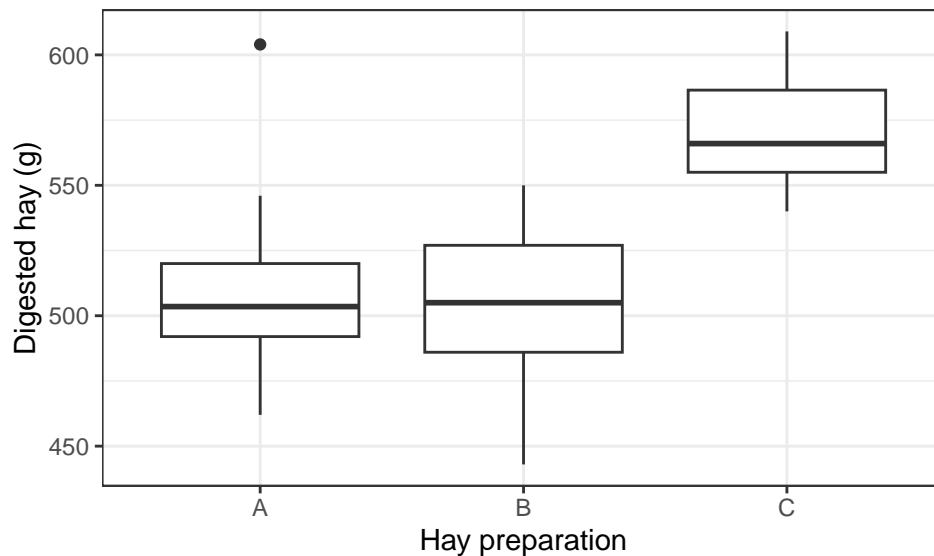
3 Latin squares

An animal scientist was conducting a feeding trial to determine the dry matter digestibility of three different preparations of hay. The scientist had only 18 sheep (experimental units) available for use in the trial, so it was decided to conduct the experiment as a series of six 3×3 Latin squares, run concurrently, with each sheep (columns = sheep) receiving each of three treatments over three consecutive feeding periods (rows = times).

```
sheep <- read.table("sheep.csv", header=T, sep="," ,
                    colClasses = c("factor","factor","factor",
                                   "factor","numeric"))

summary(sheep)
```

square	sheep	letter	time	digest	
1:9	1	: 3	A:18	1:18	Min. :443.0
2:9	10	: 3	B:18	2:18	1st Qu.:495.8
3:9	11	: 3	C:18	3:18	Median :525.0
4:9	12	: 3			Mean :528.0
5:9	13	: 3			3rd Qu.:555.0
6:9	14	: 3			Max. :609.0
	(Other):36				



3.1 aov()

```
fit_aov_LS <- aov(digest ~ square + square/sheep + square/time + letter,  
                 data=sheep)  
summary(fit_aov_LS)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
square	5	6718	1344	1.383	0.269
letter	2	50877	25438	26.180	1.52e-06 ***
square:sheep	12	7412	618	0.636	0.790
square:time	12	8852	738	0.759	0.683
Residuals	22	21377	972		

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

3.2 lm()

```
fit_lm_LS <- lm(digest ~ square + square/sheep + square/time + letter,  
               data=sheep)  
anova(fit_lm_LS)
```

Analysis of Variance Table

Response: digest

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
square	5	6718	1343.5	1.3827	0.2690
letter	2	50877	25438.4	26.1796	1.52e-06 ***
square:sheep	12	7412	617.7	0.6357	0.7901
square:time	12	8852	737.7	0.7592	0.6832
Residuals	22	21377	971.7		

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

3.3 mixed-effects model

Apparently, the variation due to sheep and time is small enough that it collapses to 0 in the variance components (see top of summary() output).


```
fit_lmer_LS <- lmer(digest ~ letter + (1|square/sheep) + (1|square:time), data=sheep)
```

boundary (singular) fit: see help('isSingular')

```
anova(fit_lmer_LS, type=3)
```

Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
letter	50877	25438	2	46	31.087	2.875e-09 ***

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

```
summary(fit_lmer_LS)
```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmerModLmerTest]

Formula: digest ~ letter + (1 | square/sheep) + (1 | square:time)

Data: sheep

REML criterion at convergence: 498

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.9434	-0.5187	-0.1436	0.6654	3.0413

Random effects:

Groups	Name	Variance	Std.Dev.
square:time	(Intercept)	0.00	0.000
sheep:square	(Intercept)	0.00	0.000
square	(Intercept)	58.36	7.639
Residual		818.29	28.606

Number of obs: 54, groups: square:time, 18; sheep:square, 18; square, 6

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	508.778	7.429	21.184	68.487	< 2e-16 ***
letterB	-4.944	9.535	46.000	-0.519	0.607
letterC	62.500	9.535	46.000	6.555	4.25e-08 ***

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

```

Correlation of Fixed Effects:
      (Intr) lettrB
letterB -0.642
letterC -0.642  0.500
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')

```

3.4 pairwise comparisons

```
emmeans(fit_lmer_LS, pairwise~"letter")
```

```

$emmeans
  letter emmean   SE    df lower.CL upper.CL
A           509 7.43 18.4     493     524
B           504 7.43 18.4     488     519
C           571 7.43 18.4     556     587

```

```

Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95

```

```

$contrasts
 contrast estimate   SE df t.ratio p.value
A - B           4.94 9.54 22   0.519  0.8632
A - C          -62.50 9.54 22  -6.555 <.0001
B - C          -67.44 9.54 22  -7.073 <.0001

```

```

Degrees-of-freedom method: kenward-roger
P value adjustment: tukey method for comparing a family of 3 estimates

```

4 R Activity

You will want to ensure that you use Type III sums of squares when conducting ANOVAs (i.e., use marginal and not sequential fits). You will need the following packages to complete this problem set:

```
library(tidyverse)
library(car)
library(lme4)
library(lmerTest)
library(emmeans)
```

A turfgrass management study was initiated in to determine if tillage plan and herbicide applications to control quackgrass (a noxious weed) influenced seed production in a single, commonly grown variety of perennial ryegrass. The experiment was set up as a 3 x 2 factorial with 3 replications (=blocks) in a randomized complete block design (ignore interactions for this problem set, but note that factorial designs are often implemented to evaluate the interactive effect of two treatments on a single response variable). The data are in the “EPP_seed.txt” data file.

- Factor 1: Tillage, 3 kinds
 - Spring = spring tillage/spring seeding
 - Fall = fall tillage/fall seeding
 - None = no-tillage/fall seeding
- Factor 2: Herbicide, 2 levels
 - None = no herbicide treatment
 - Applied = herbicide treatment (to control quackgrass)

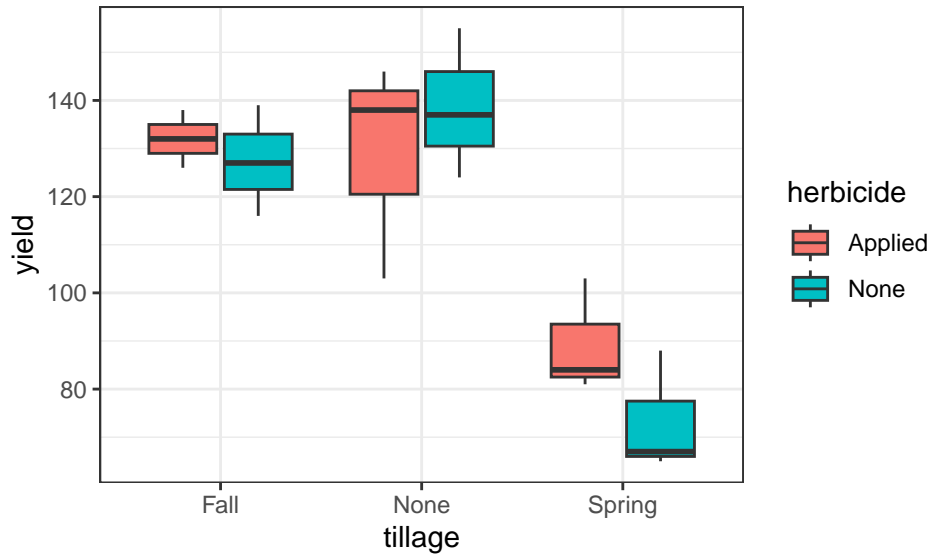
1. Load in the data. Note that `rep` is the column name for blocks.

```
seed_df <- read.table('EPP_seed.txt', header=T, sep="\t",
                      colClasses = c("factor", "factor", "factor", NA))
summary(seed_df)
```

rep	tillage	herbicide	yield
1:6	Fall :6	Applied:9	Min. : 65.00
2:6	None :6	None :9	1st Qu.: 91.75
3:6	Spring:6		Median :125.00
			Mean :114.94
			3rd Qu.:137.75
			Max. :155.00

2. Graph the data using a boxplot. In the plot, group the data by tillage treatment on the x-axis and then color each box by herbicide treatment.

```
ggplot(seed_df, aes(x=tillage, y=yield, fill=herbicide)) +
  geom_boxplot() + theme_bw()
```



- Conduct an analysis of variance (ANOVA) using the `lm()` and `Anova()` (from the `car` package) commands (i.e., assess if `tillage` and `herbicide` explain variation in `yield`)

```
lm_seed_1 <- lm(yield ~ rep + tillage + herbicide, data=seed_df)
Anova(lm_seed_1, type="III")
```

Anova Table (Type III tests)

Response: yield

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	50613	1	309.9481	6.148e-10	***
rep	1013	2	3.1031	0.08199	.
tillage	10219	2	31.2911	1.735e-05	***
herbicide	60	1	0.3705	0.55408	
Residuals	1960	12			

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

- Select a variable you would like to explore further using pairwise comparisons. Explain your reasoning, and then conduct the comparisons.

```
emmeans(lm_seed_1, pairwise~tillage)
```

\$emmeans

tillage	emmean	SE	df	lower.CL	upper.CL
Fall	129.7	5.22	12	118	141.0
None	133.8	5.22	12	122	145.2

Spring	81.3	5.22	12	70	92.7
--------	------	------	----	----	------

Results are averaged over the levels of: rep, herbicide

Confidence level used: 0.95

\$contrasts

contrast	estimate	SE	df	t.ratio	p.value
Fall - None	-4.17	7.38	12	-0.565	0.8410
Fall - Spring	48.33	7.38	12	6.551	0.0001
None - Spring	52.50	7.38	12	7.116	<.0001

Results are averaged over the levels of: rep, herbicide

P value adjustment: tukey method for comparing a family of 3 estimates

- Write 3-4 sentences summarizing your findings using “biologically meaningful” terms.

Answer: Yield of the perennial rye grass varied significantly between tillage treatments ($F_{1,12} = 31.29, p < 0.0001$) but herbicide had no effect ($F_{1,12} = 0.37, p = 0.55$), indicating that quackgrass is best managed by altering tillage. A pairwise comparison between tillage options indicated that tilling/seeding in spring significantly reduced yield by approx. 48g and 53g compared with tilling/seeding in fall ($t_{12} = 6.55, p = 0.0001$) and no tilling/fall seeding ($t_{12} = 7.12, p < 0.0001$). Treatments using fall seeding did not differ ($t_{12} = -0.57, p = 0.84$). Taken together, tillage and seeding should be completed in fall.