Statistical Modelling: Understanding Variance/Error Structure

Chapter 4

Terry Neeman and Timothée Bonnet

November 21, 2018

Statistical models: MEAN and VARIANCE components

$$\textit{response} = \underbrace{A + D \times \textit{predictor}}_{\substack{\text{Mean Structure} \\ \text{Experimental factors}}} + \underbrace{\epsilon, \text{ with } \epsilon \sim N(0, \sigma)}_{\substack{\text{Variance Structure} \\ \text{Unrelated to experiment factors} \\ \text{Unexplained ``noise''}}}$$

Statistical models: MEAN and VARIANCE components

$$response = \underbrace{A + D \times predictor}_{\text{Mean Structure}} + \underbrace{\epsilon, \text{ with } \epsilon \sim N(0, \sigma)}_{\text{Variance Structure}}$$

Mean Structure
Experimental factors

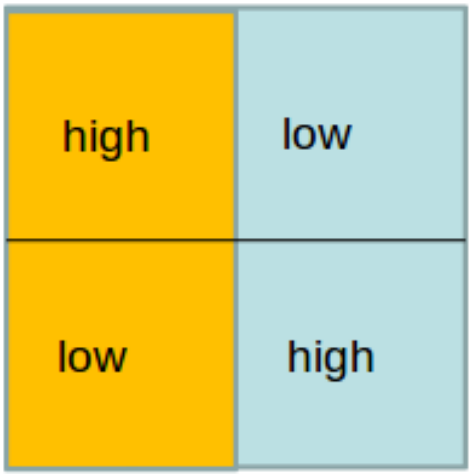
Variance Structure
Unrelated to experiment factors
Unexplained "noise"

What is in ϵ? How can we tweak that? Why should we care?

Describe the data structure in this experiment

How is seedling emergence (in Banksia) influenced by temperature and moisture?

Shed 1



Shed 2

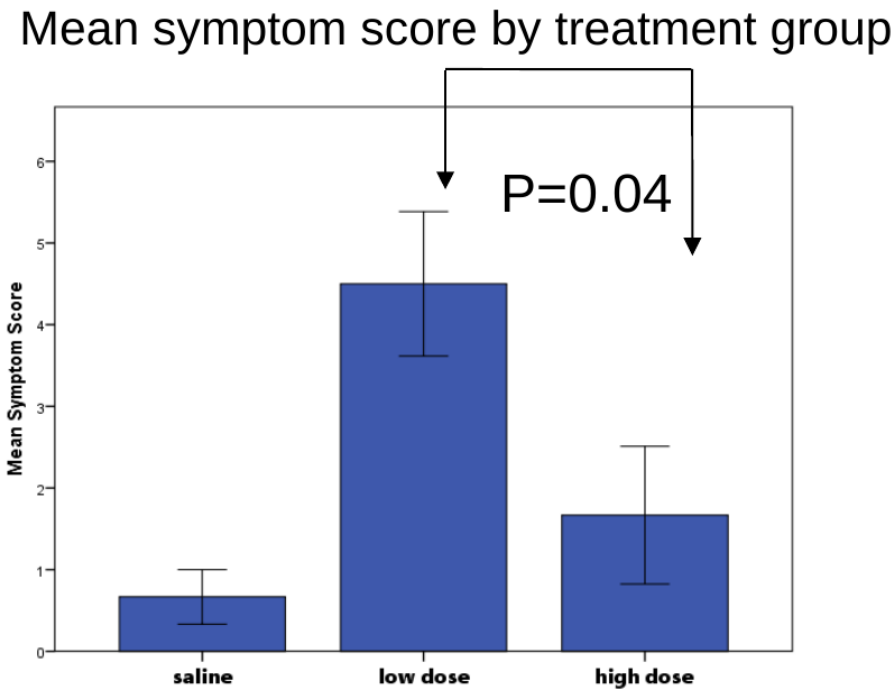
Shed 3

Set up: 3 sheds, 4 garden beds per shed, 24 pots per bed.
Experimental factors: 2 temperatures, 2 water levels

Key components of a statistical model of an experiment

- Outcome measure
  - Plant height at week 3
  - Number of leaves at week 3
- Experimental factors
  - Temperature (warm/ control)
  - Watering conditions (low/high)
  - Question: how does each factor affect outcome measures? Do the factors interact?
- **Blocking factors**
  - Shed
  - "half-shed" within Shed
  - Garden bed within "half-shed"

Message 1: A small p-value is not always evidence of a treatment effect



Vaccine challenge experiment:

- 6 mice/group (saline/low dose/high dose)
- All mice challenged with Shigella
- Followed for 14 days
- Outcome: Symptom score average Days 2 - 8

One-way ANOVA (post-hoc Bonferroni) p=0.04

Noise confounded with treatment

Experimental design

The observed difference in outcome could be the result of:

- Cage effects
- Mouse strain effects

These effects are CONFOUNDED with treatment effect

Cage 1:
saline

Cage 2:
Low Dose

Cage 3:
High Dose

Noise confounded with treatment

Experimental design

The observed difference in outcome could be the result of:

- Cage effects
- Mouse strain effects

These effects are CONFOUNDED with treatment effect

Cage 1:
saline

Cage 2:
Low Dose

Cage 3:
High Dose

Solutions:

Mixed cages: can compare within cages

More cages: must compare between cages

Noise confounded with treatment

Mixed cages: can compare within cages

- **Share the noise among treatments**
- Few cages needed: Technically efficient
- But may be technically impossible

Noise confounded with treatment

Mixed cages: can compare within cages

- **Share the noise among treatments**
- Few cages needed: Technically efficient
- But may be technically impossible

More cages: must compare between cages

- **Redefine experimental unit**
- Noise among cages, instead of within
- Needs to re-scale the experiment

Is photosynthetic rate affected by temperature?

Research context

- Outcome measure: Photosynthetic rate
- Experimental factors: Temperature (high/low)
- Blocking factors: Position (4)

How many parameters?

Is photosynthetic rate affected by temperature?

Research context

- Outcome measure: Photosynthetic rate
- Experimental factors: Temperature (high/low)
- Blocking factors: Position (4)

How many parameters?

2 parameters to describe the effect of temperatures

Is photosynthetic rate affected by temperature?

Research context

- Outcome measure: Photosynthetic rate
- Experimental factors: Temperature (high/low)
- Blocking factors: Position (4)

How many parameters?

2 parameters to describe the effect of temperatures

+ some to correct for blocking factors

Inference using linear model without and with blocking in R

Without blocking

```
m_noblock <- lm(PhotoRate~Temp, data=photo)
anova(m_noblock)
```

```
Analysis of Variance Table

Response: PhotoRate
           Df  Sum Sq  Mean Sq  F value  Pr(>F)
Temp        1   613.3   613.28   0.4386  0.5147
Residuals  22 30760.8  1398.22
```

Inference using linear model without and with blocking in R

Blocking with a fixed factor

```
m_block <- lm(PhotoRate~Temp+ as.factor(Position), data=photo)
anova(m_block)
```

```
Analysis of Variance Table

Response: PhotoRate
                   Df  Sum Sq Mean Sq F value    Pr(>F)
Temp                1   613.3   613.3   4.521   0.04681 *
as.factor(Position)  3 28183.4  9394.5  69.253 2.047e-10 ***
Residuals           19  2577.4   135.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Clearer evidence for an effect of temperature with blocking

Inference using linear model without and with blocking in R

Blocking with a **random effect**

```
library(lme4)
library(lmerTest)
m_block_re <- lmer(PhotoRate~Temp+ (1|Position), data=photo)
anova(m_block_re)
```

```
Type III Analysis of Variance Table with Satterthwaite's metho
     Sum Sq Mean Sq NumDF DenDF F value  Pr(>F)
Temp 613.28  613.28     1    19   4.521 0.04681 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearer evidence for an effect of temperature with blocking

Try it in R

1. Load data "Prac4photosynthesis.csv"
2. Visualize the data
3. Model data and interpret output

```
library(lmerTest)
library(emmeans)
lmer1<-lmer(PhotoRate~Temp+(1|Position), data=photo)
anova(lmer1)
summary(lmer1)
emmeans(lmer1,~Temp)
```

4. Assess model assumptions

```
plot(lmer1)
```

Fixed or random effect?

In this example

- Doesn't change inference (same p-value for temperature)
- Summary cleaner with random effect

Fixed or random effect?

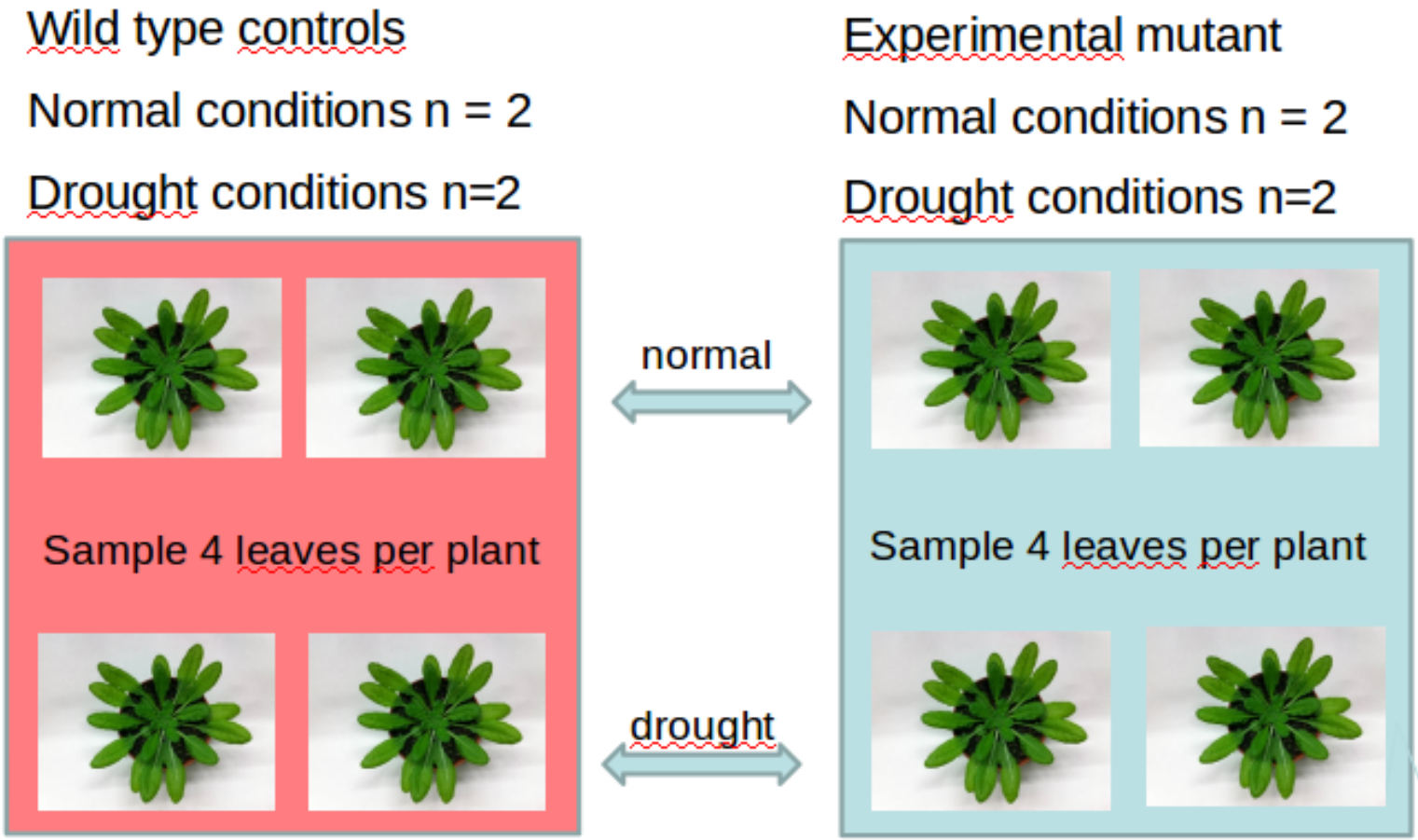
In this example

- Doesn't change inference (same p-value for temperature)
- Summary cleaner with random effect

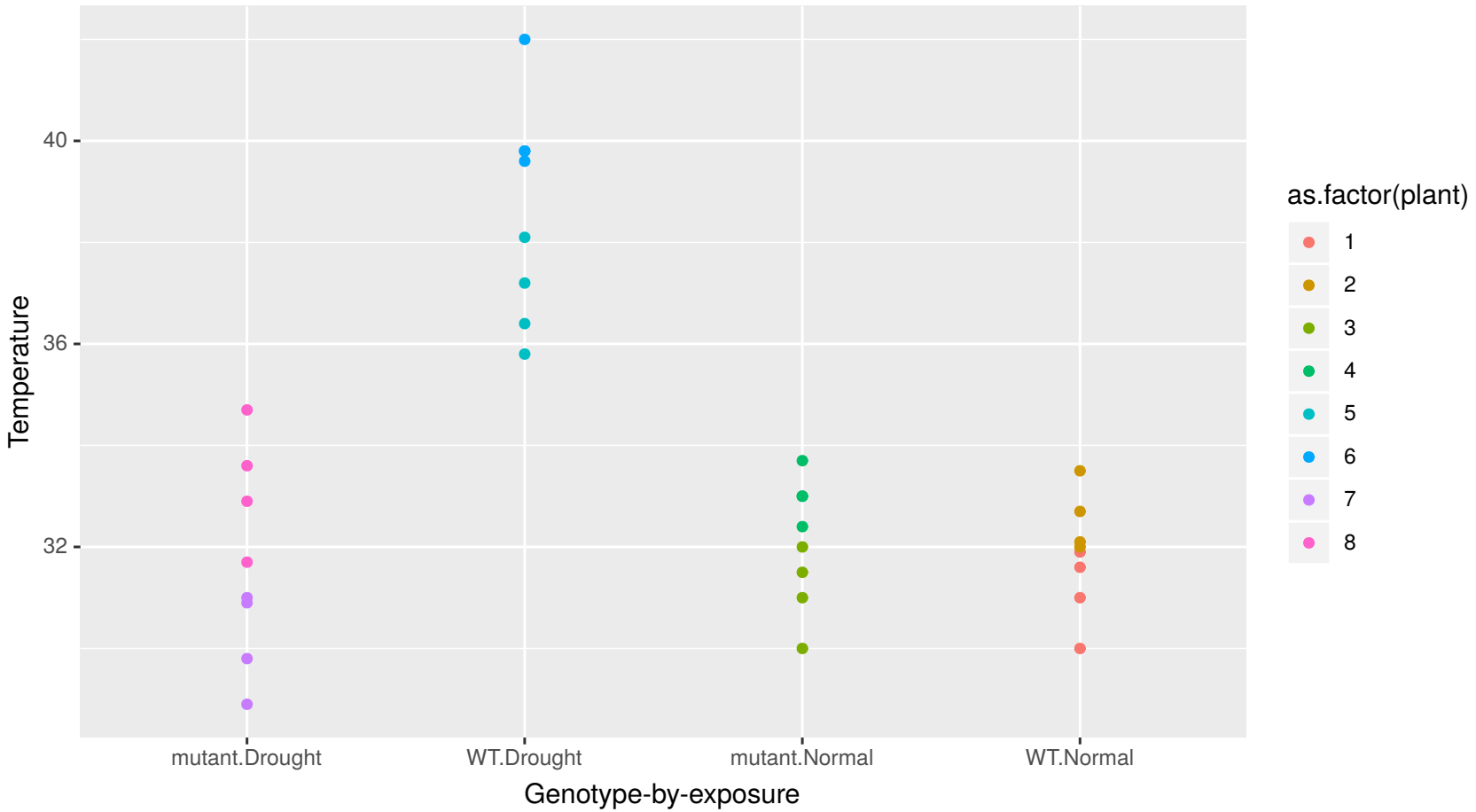
In general

- Generally doesn't change inference much. Random effect slightly more efficient.
- Summary cleaner with random effect, especially when many random levels
- Random shifts the focus from level values to variation among levels
- Variance parameters interesting in themselves
- Are levels of interest (fixed) or are they some kind of noise (random)

Can a gene KO Arabidopsis modulate leaf temperature during drought?



Visualising temperature data by treatment and Genotype



Outcome measure: Temperature

Experimental factors: Genotype (2), Watering conditions (2)

Blocking factor: Plant

Set up analysis for this experiment

```
drought <- read.csv("Data/Prac3droughtdata.csv")
str(drought)

#Make Plant a Factor
drought$plant<- factor(drought$plant)

#Set Reference Levels
drought$Genotype<-relevel(drought$Genotype, ref="WT")
drought$WaterCondition<-relevel(drought$WaterCondition, ref="Normal")

ggplot(drought, aes(x=interaction(Genotype, WaterCondition),
     y=Temperature, color=plant))+
  geom_point()+xlab("Genotype-by-exposure")
```

Analysis of Variance without and with variance structure in R

Analysis of Variance without and with variance structure in R

```
 lm.drought <- lm(Temperature ~ Genotype*WaterCondition, data=drought)
anova(lm.drought)
```

Analysis of Variance without and with variance structure in R

```
 lm.drought <- lm(Temperature ~ Genotype*WaterCondition, data=drought)
anova(lm.drought)
```

```
                          Df  Sum Sq Mean Sq F value    Pr(>F)
Genotype                   1  89.111  89.111  33.289 3.407e-06 ***
WaterCondition             1  80.645  80.645  30.127 7.304e-06 ***
Genotype:WaterCondition    1 101.531 101.531  37.929 1.195e-06 ***
```

Analysis of Variance without and with variance structure in R

```
 lm.drought <- lm(Temperature ~ Genotype*WaterCondition, data=drought)
anova(lm.drought)
```

```
                          Df  Sum Sq Mean Sq F value    Pr(>F)
Genotype                   1  89.111  89.111  33.289 3.407e-06 ***
WaterCondition             1  80.645  80.645  30.127 7.304e-06 ***
Genotype:WaterCondition    1 101.531 101.531  37.929 1.195e-06 ***
```

```
lmer.drought <- lmer(Temperature ~ Genotype*WaterCondition + (1|plant),
data=drought)
anova(lmer.drought)
```

```
                          Sum Sq Mean Sq NumDF DenDF F value  Pr(>F)
Genotype                  5.8403  5.8403     1     4  6.6257 0.06171 .
WaterCondition            5.2854  5.2854     1     4  5.9962 0.07054 .
Genotype:WaterCondition   6.6543  6.6543     1     4  7.5491 0.05150 .
```

Treatment effect estimates without and with variance structure

```
emmeans(lm.drought,~Genotype*WaterCondition)
```

```
Genotype Condition  emmean    SE  df  lower.CL upper.CL
WT       Normal      31.85  0.578 28  30.66  33.03
WT       Drought     38.58  0.578 28  37.40  39.77
mutant   Normal      32.07  0.578 28  30.89  33.25
mutant   Drought     31.68  0.578 28  30.50  32.87
```

```
emmeans(lmer.drought,~Genotype*WaterCondition)
```

```
Genotype Condition emmean     SE  df   lower.CL upper.CL
WT   Normal    31.85     1.29  4    28.25  35.44
WT   Drought   38.58     1.29  4    34.98  42.18
mutant   Normal    32.07     1.29  4    28.47  35.67
mutant   Drought   31.68     1.29  4    28.08  35.28
```

Correct blocking structure is essential for correct inference!

Is dark respiration differentially affected by temperature between genotypes?

Research context

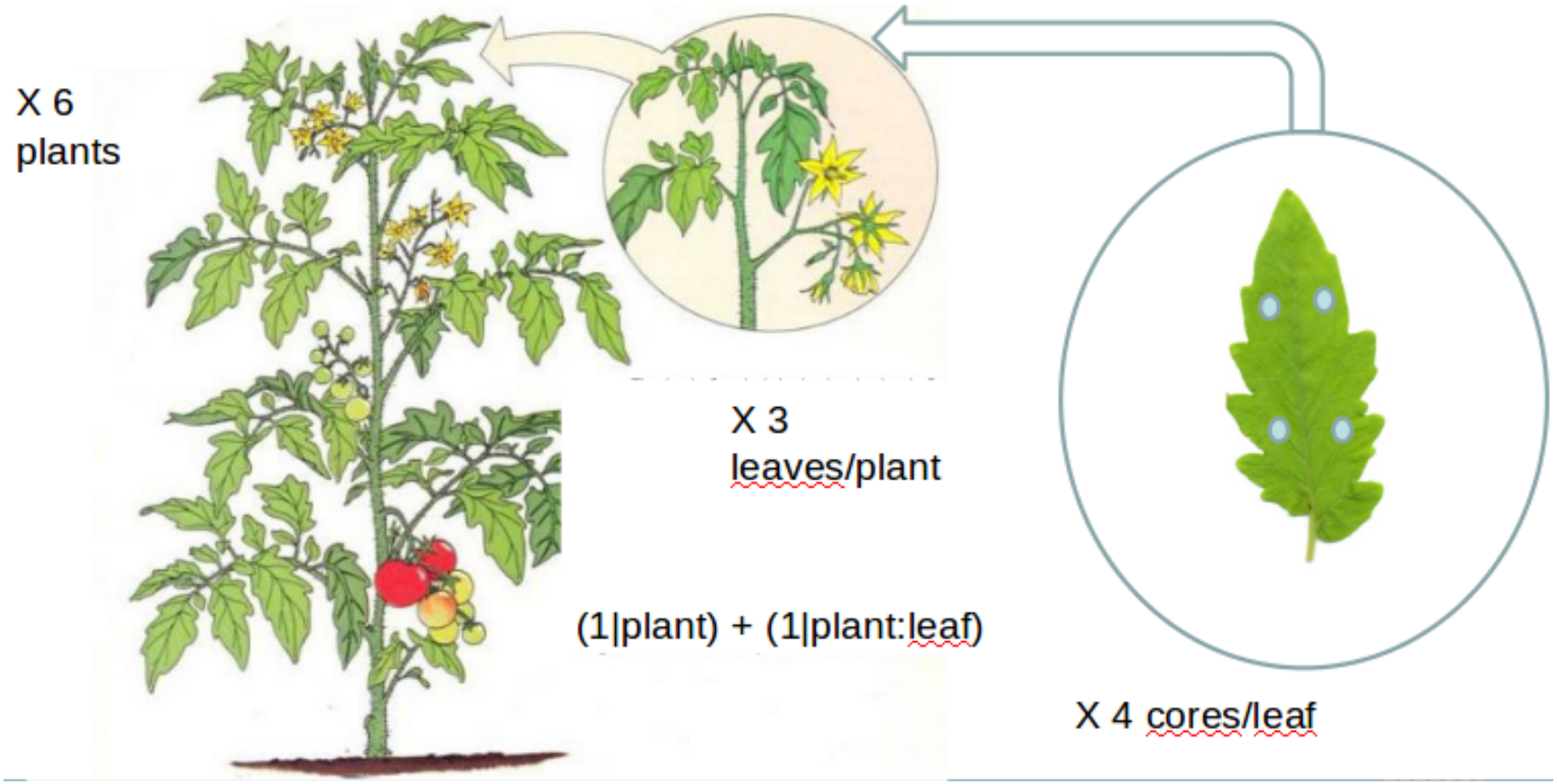
- Outcome measure: Dark respiration
- Experimental factors: Genotype (2) & Temperature (4)
- Blocking factors: Shelter (4) & Plants within shelter (20)

8 parameter model (plus random effects)

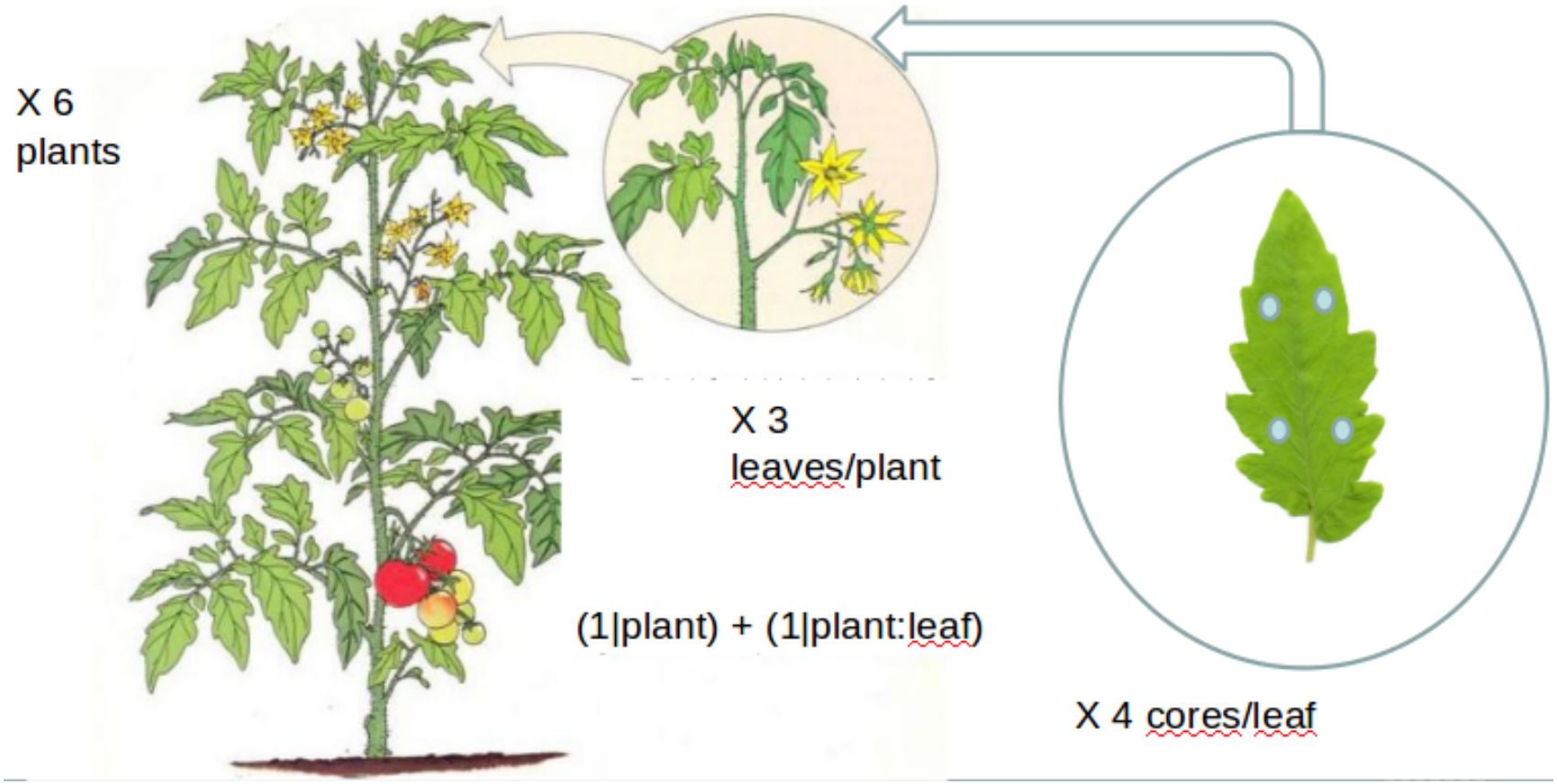
Is dark respiration differentially affected by temperature between genotypes?

Analyse the data "Prac4darkrespiration.csv"
Answer the question
Check assumptions
Plot result

Understanding different variance structure



Understanding different variance structure



Understanding different variance structure

X 6 cages

X 4 time points/mouse

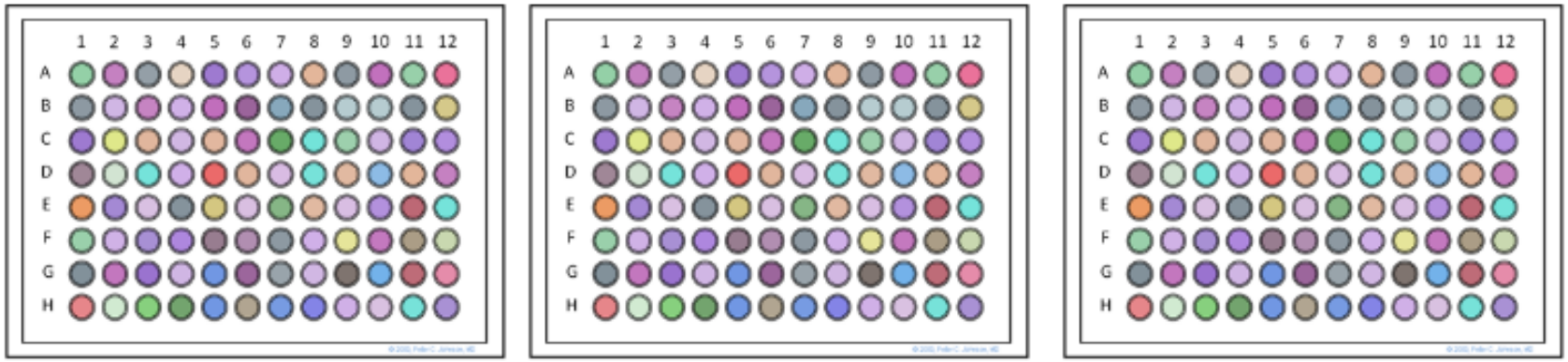
(1|cage) + (1|cage:mouse)

Understanding different variance structure: **Nested and Crossed structures**

Crossed: `(1|plate) + (1|row) + (1|column)`

Nested: `(1|plate) + (1|plate:row) + (1|plate:column) = (1|plate/row/column)`

What is the difference?



crossed random effects: one level of a random effect can appear in conjunction with more than one level of another random effect

Everything you need to know about mixed models

- `http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html`
- Subscribe to mailing-list:
  `https://stat.ethz.ch/mailman/listinfo/r-sig-mixed-models`

Summary of Statistical Modelling

Take-home

- Identify Statistical Framework of Experiment

Summary of Statistical Modelling

Take-home

- Identify Statistical Framework of Experiment
  1. Outcome measure

Summary of Statistical Modelling

Take-home

- Identify Statistical Framework of Experiment
  1. Outcome measure
  2. Experimental factors

Summary of Statistical Modelling

Take-home

- Identify Statistical Framework of Experiment
  1. Outcome measure
  2. Experimental factors
  3. Blocking factors

Summary of Statistical Modelling

Take-home

- Identify Statistical Framework of Experiment
  1. Outcome measure
  2. Experimental factors
  3. Blocking factors
- Visualize data

Summary of Statistical Modelling

Take-home

- Identify Statistical Framework of Experiment
  1. Outcome measure
  2. Experimental factors
  3. Blocking factors
- Visualize data
- Try simple models first

Summary of Statistical Modelling

Take-home

- Identify Statistical Framework of Experiment
  1. Outcome measure
  2. Experimental factors
  3. Blocking factors
- Visualize data
- Try simple models first
- Assess model fit/assumptions

Summary of Statistical Modelling

Take-home

- Identify Statistical Framework of Experiment
  1. Outcome measure
  2. Experimental factors
  3. Blocking factors
- Visualize data
- Try simple models first
- Assess model fit/assumptions
- interpret