# Statisitcal Thinking in Biology Research

Terry Neeman and Timothee Bonnet

November 20, 2018

# Acknowledgemnts and warning

# Key ideas for today

- Statistics in biology is the study of biological variation

# Key ideas for today

- Statistics in biology is the study of biological variation
- Statistical ideas about biological variation inform the design of experiments

# Key ideas for today

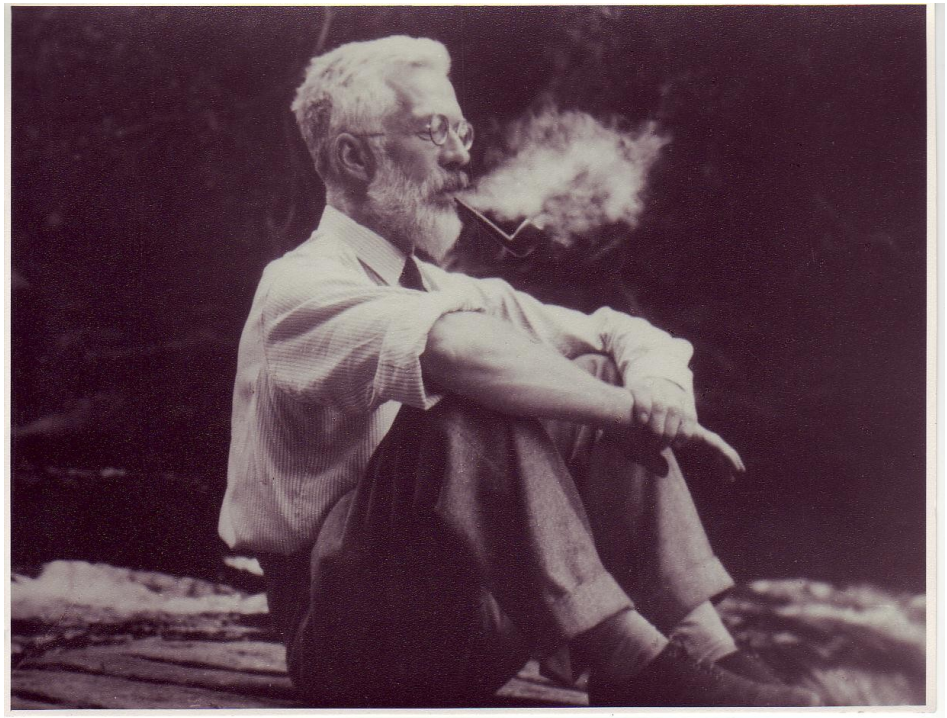
- Statistics in biology is the study of biological variation
- Statistical ideas about biological variation inform the design of experiments
- Statistical ideas about biological variation inform the analysis of experiments

# Key ideas for today

- Statistics in biology is the study of biological variation
- Statistical ideas about biological variation inform the design of experiments
- Statistical ideas about biological variation inform the analysis of experiments
- Statistical thinking is an essential component of scientific thinking

# A bit of history of statistical methods

R.A. Fisher: 1890-1962



Statistical Principles for Research Workers (1925)

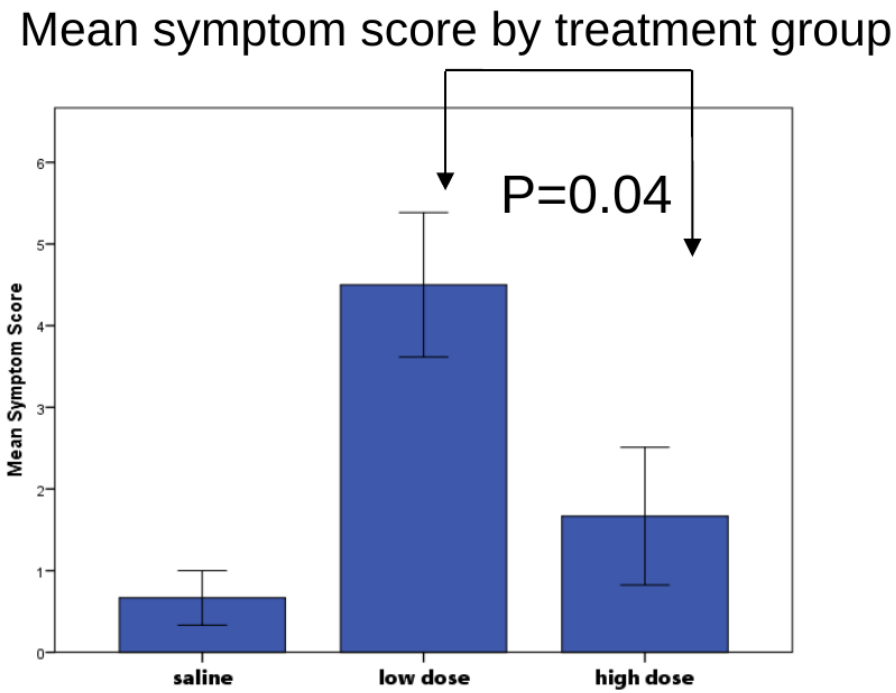
# A bit of history of statistical methods

R.A. Fisher: 1890-1962

Statistical Principles for Research Workers (1925)

1. Cautionary tales from the front
2. Introduction to Statistical Modelling
3. Another look at essential steps

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

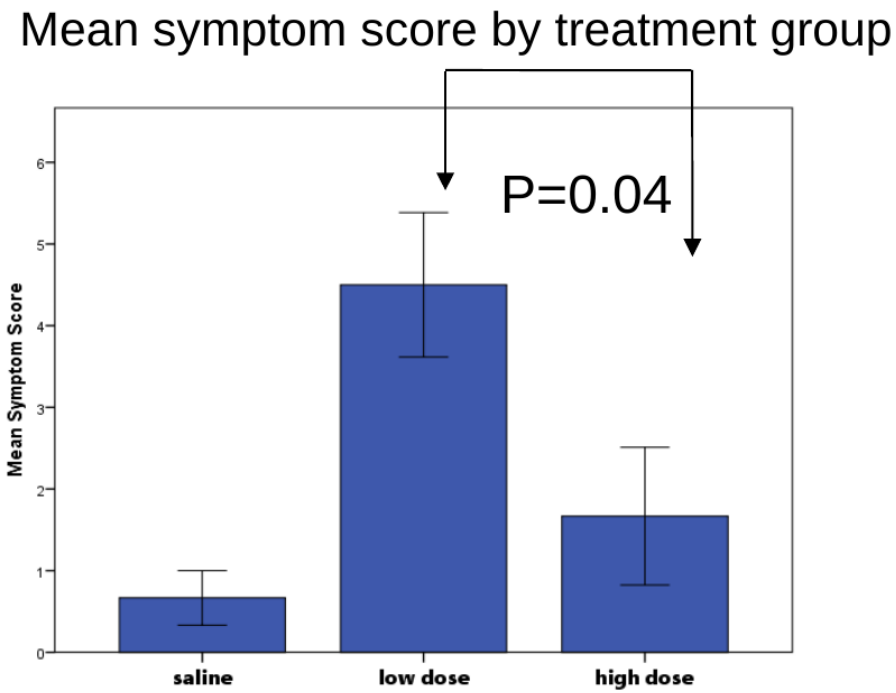



## Vaccine challenge experiment:

- 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

# 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

*Do you think the vaccine works? What is strange?*

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

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

**Experimental design**

The observed difference in outcome could be the result of:

- Cage effects
- Mouse strain effects

These effects are CONFOUNDED with treatment effect

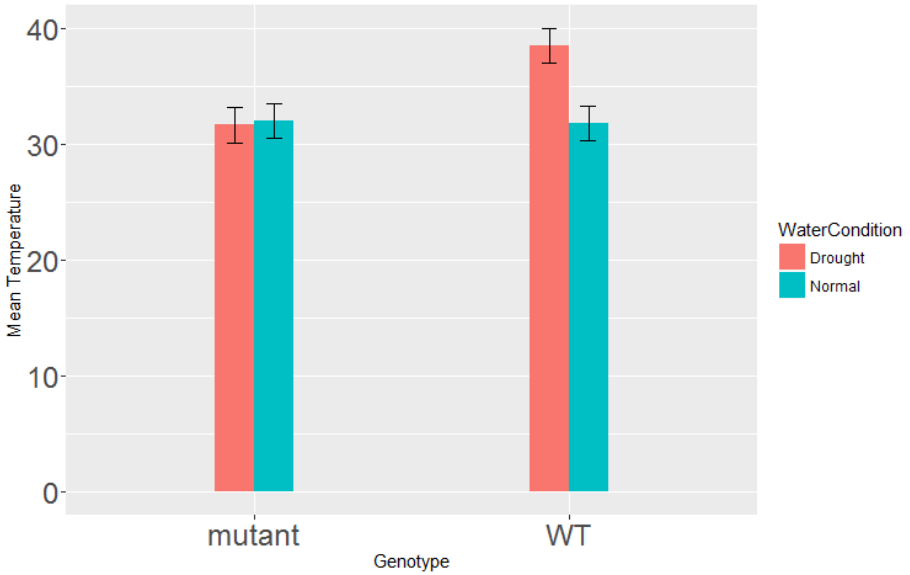
Cage 1:
saline

Cage 2:
Low Dose

Cage 3:
High Dose

# Message 2: p-values from simple comparisons cannot tell us when differences are “different”

# Message 2: p-values from simple comparisons cannot tell us when differences are "different"




**Are temperature mechanisms modified in a genetically modified tomato plant?**

- Genotypes: WT/mutant
- Water condition: Normal/Drought
- Leaf temperature measured

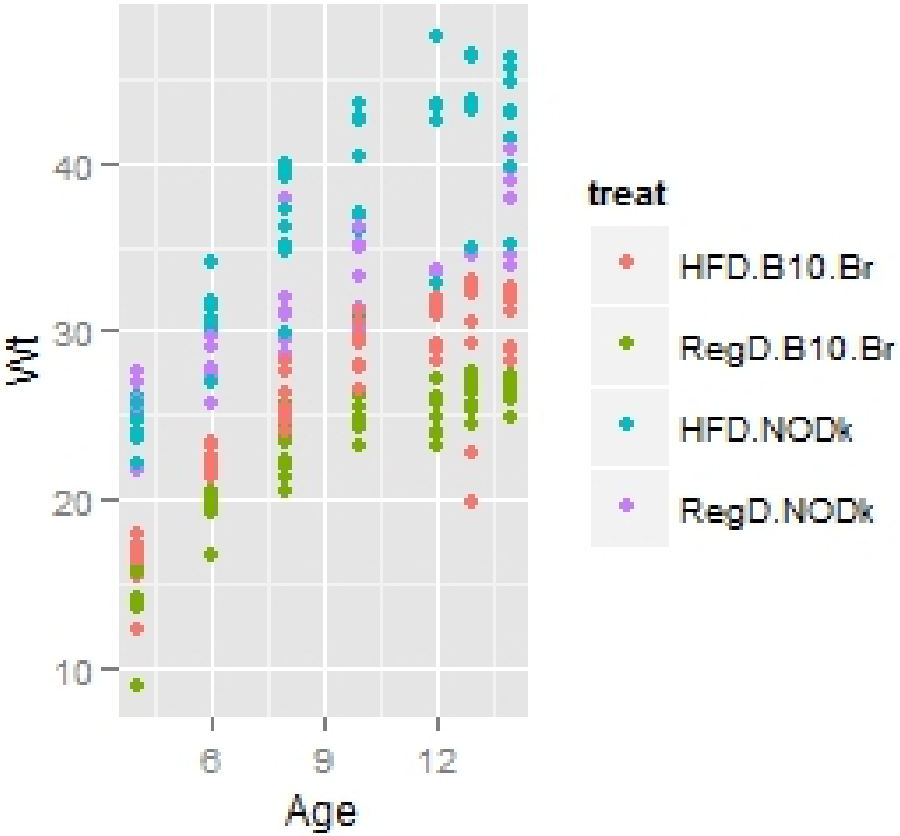
**Comparisons made using t-tests**

Evidence of difference + No evidence of difference $\neq$ Evidence that differences are different.

# Message 3: Interpreting experimental results needs more than t-tests

# Message 3: Interpreting experimental results needs more than t-tests

Research question: Are mice susceptible to obesity when exposed to a high fat diet?




## Experimental set-up:

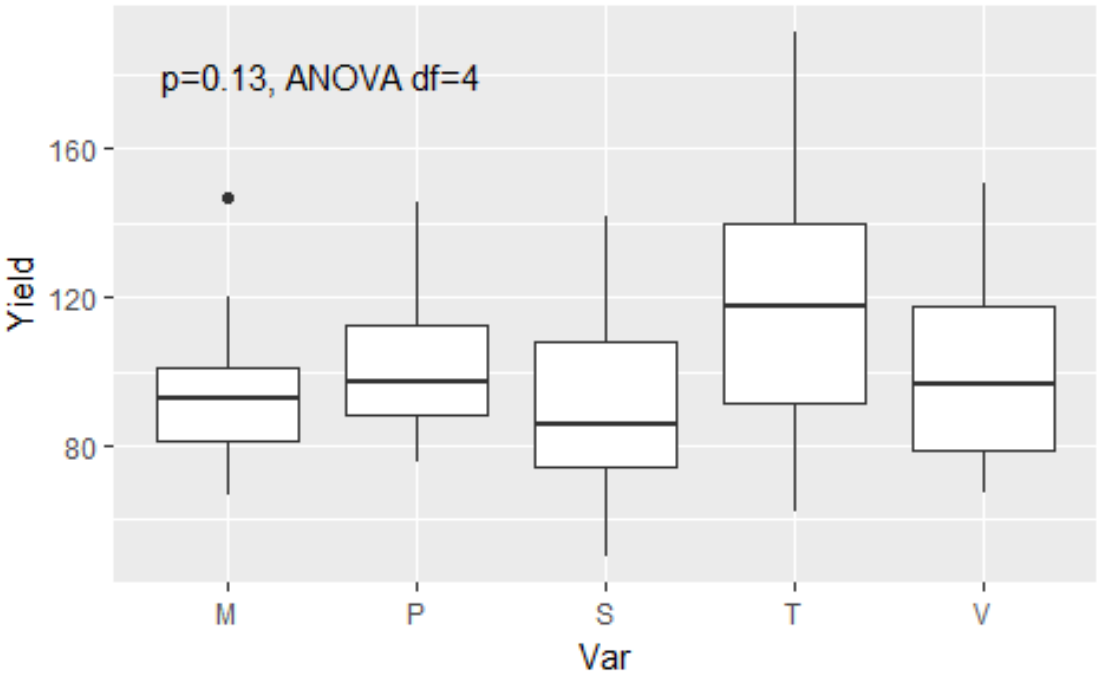
- 37 mice: 16 NODk /21 WT
- Randomised to either regular or high fat diet
- Monitored for 14 weeks
- Outcome measure: Body weight (g)
- Experimental factors: Diet (2), Strain (2), Time (8)

Acknowledgements: Ainy Hussain, PhD student 2013

# Message 4: Knowing how to combine information across subgroups can improve inference

# Message 4: Knowing how to combine information across subgroups can improve inference

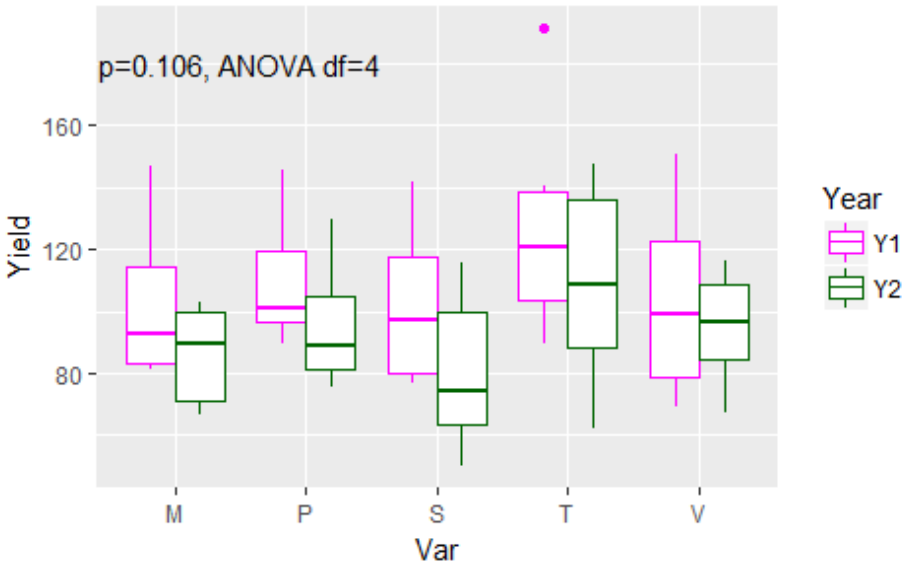
Comparing yield in five barley varieties (1930s)
Experimental factors: 5 varieties of barley, 6 locations, 2 time points.
Outcome measure: yield




Acknowledgements: MASS R-package

# Message 4: Knowing how to combine information across subgroups can improve inference

Comparing yield in five barley varieties (1930s)
Experimental factors: 5 varieties of barley, 6 locations, 2 time points.
Outcome measure: yield



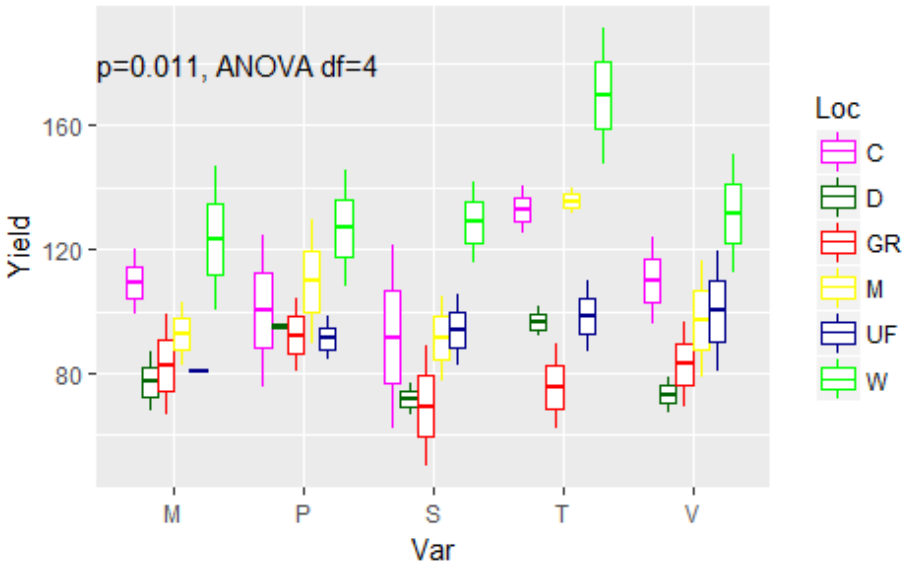

## Controlling for other sources of variation:

- Controlling for year = comparing yield WITHIN years and combining these

Acknowledgements: MASS R-package

# Message 4: Knowing how to combine information across subgroups can improve inference

Comparing yield in five barley varieties (1930s)
Experimental factors: 5 varieties of barley, 6 locations, 2 time points.
Outcome measure: yield

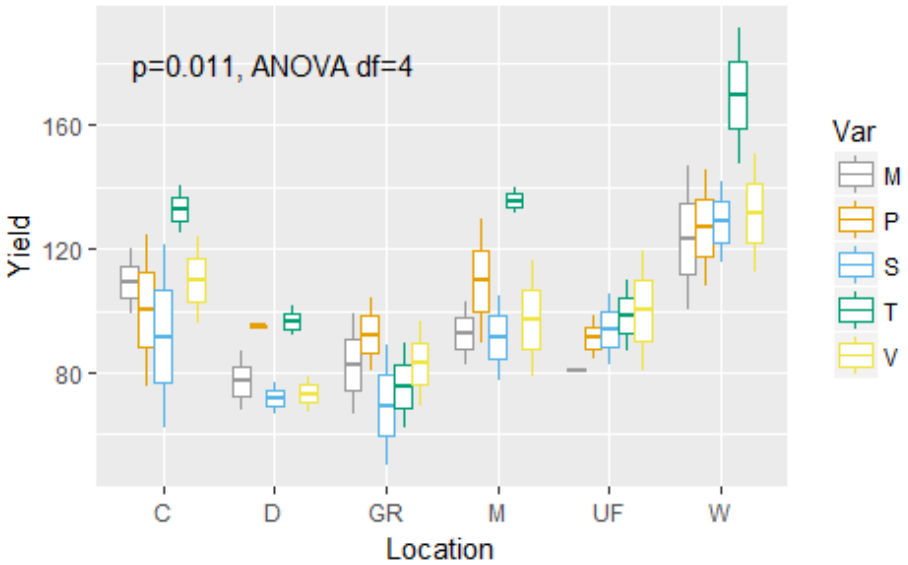



## Controlling for other sources of variation:

- Control for year = compare yield WITHIN years and combine these
- Control for location = compare yield WITHIN locations and combine these

Acknowledgements: MASS R-package

# Message 4: Knowing how to combine information across subgroups can improve inference




## Controlling for other sources of variation:

- Control for year = compare yield WITHIN years and combine these
- Control for location = compare yield WITHIN locations and combine these

Acknowledgements: MASS R-package

# Message 5: Knowing what factors contribute to the variation in outcome helps design experiments and analyses

# Message 5: Knowing what factors contribute to the variation in outcome helps design experiments and analyses

Research question: How does cold duration impact upon germination in alpine plant A. glacialis?

## Experimental set-up:

- Seed collections from alpine region in Australia
- 3 Regions- low/high altitude
- 4 sets of Petri dishes
- 4 cabinet shelves
- Response - % germinated

**What other factors are important to consider when comparing cold duration?**

# Summary

1. A small p-value is not always evidence of a treatment effect. **Good experimental design matters.**

# Summary

1. A small p-value is not always evidence of a treatment effect. **Good experimental design matters.**
2. p-values from simple comparisons cannot tell us when differences are "different". **For each question / comparison, a specific test**

# Summary

1. A small p-value is not always evidence of a treatment effect. **Good experimental design matters.**
2. p-values from simple comparisons cannot tell us when differences are "different". **For each question / comparison, a specific test**
3. Interpreting experimental results needs more than t-tests. **Need a statistical model of the experiment, matching scientific question.**

# Summary

1. A small p-value is not always evidence of a treatment effect. **Good experimental design matters.**
2. p-values from simple comparisons cannot tell us when differences are "different". **For each question / comparison, a specific test**
3. Interpreting experimental results needs more than t-tests. **Need a statistical model of the experiment, matching scientific question.**
4. Combining information across subgroups can improve inference. **A statistical model enables accumulation of evidence across experiments.**

# Summary

1. A small p-value is not always evidence of a treatment effect. **Good experimental design matters.**
2. p-values from simple comparisons cannot tell us when differences are "different". **For each question / comparison, a specific test**
3. Interpreting experimental results needs more than t-tests. **Need a statistical model of the experiment, matching scientific question.**
4. Combining information across subgroups can improve inference. **A statistical model enables accumulation of evidence across experiments.**
5. Knowing what factors contribute to the variation in outcome matters. **A statistical model allows one to incorporate effect of other factors in the analysis.**

1. Cautionary tales from the front
2. Introduction to Statistical Modelling
3. Another look at essential steps

# Introduction to Statistical Modelling

- What is a statistical model?
- Modelling outcomes:
  - a summary of data
  - a prediction model
  - an explanatory model
- Model – may take many different functional forms
- Model – a conceptualization of the experiment

# Introduction to Statistical Modelling

- What is a statistical model?
- Modelling outcomes:
  - a summary of data
  - a prediction model
  - an explanatory model
- Model – may take many different functional forms
- Model – a conceptualization of the experiment

ALWAYS BEGIN WITH A RESEARCH QUESTION

# Key components of a statistical model of an experiment

- Outcome measure
  - Response variable
  - Measure of interest
- Experimental factors
  - Conditions that can be manipulated
  - Conditions of interest (e.g. genotype, gender)
  - Main questions: do the conditions impact upon the outcome measure?
- Blocking factors
  - Conditions (not of interest) that may impact upon the outcome measure
  - Sources of variation in the experiment that need to be controlled for
  - Clustering of experimental units

ALWAYS BEGIN WITH A RESEARCH QUESTION

# Key Objectives of a statistical model of an experiment

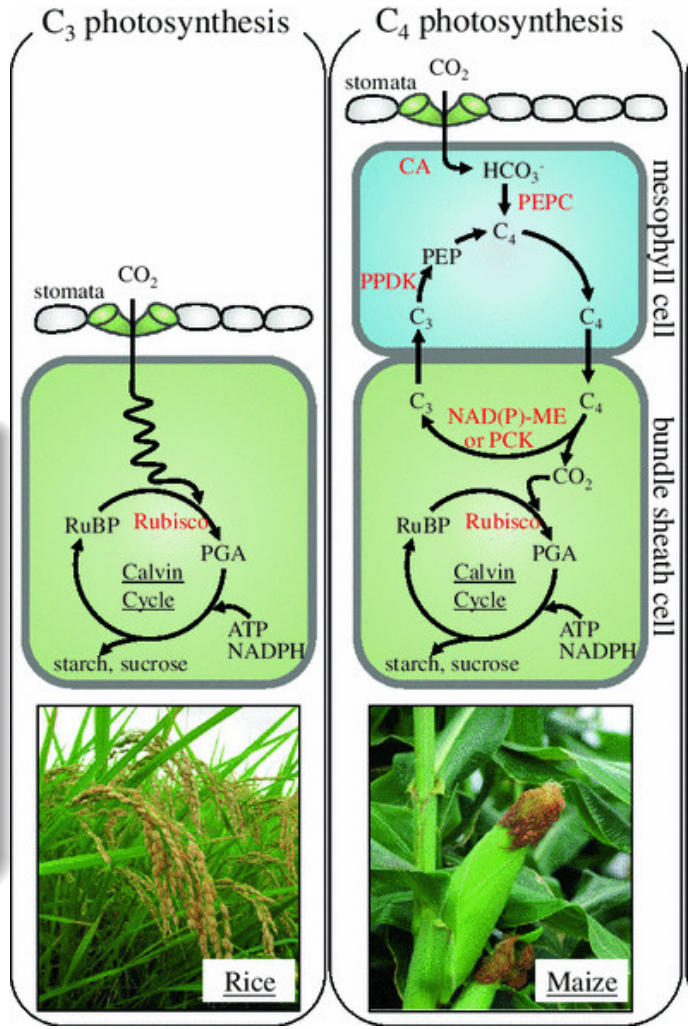
- To compare the mean response of an organism/system to a set of different experimental conditions.
  - Obtain estimate of "Treatment effect"
  - Is this "effect" different in subgroups of interest?
- What are the most important factors influencing the mean response?
- Subsidiary question: how can we design our experiment in future to more efficiently test our hypotheses?

# Example 1: Does dark respiration differ between C3 and C4 plants?

Outcome measure: dark respiration
Experimental factor: Plant type (C4/C3)
Data: 6 plants each of C4, C3

## Can calculate

- Observed overall mean
- Observed mean C3 plants
- Observed mean C4 plants
- Variation around each mean



# Example 1: Does dark respiration differ between C3 and C4 plants?

## Can calculate

- Observed overall mean
- Observed mean C3 plants
- Observed mean C4 plants
- Variation around each mean

**Statistical model**

Respiration = Mean for C3 + Difference C4-C3 * (is C4?) + Noise

# Example 1: Does dark respiration differ between C3 and C4 plants?

## Can calculate

- Observed overall mean
- Observed mean C3 plants
- Observed mean C4 plants
- Variation around each mean

**Statistical model**

Respiration = Mean for C3 + Difference C4-C3 * (is C4?) + Noise

$response = A + D \times predictor + \epsilon$

$A$ and $D$ are the model PARAMETERS.

We want to infer whether $D$ is different from 0

# Example 1: Does dark respiration differ between C3 and C4 plants?

$response = A + D \times predictor + \epsilon$

Can we separate the signal $D$ from the noise $\epsilon$ ?

# Example 1: Does dark respiration differ between C3 and C4 plants?

$response = A + D \times predictor + \epsilon$

Can we separate the signal $D$ from the noise $\epsilon$ ?

## T-test

- Outcome is a continuous variable
- Experimental factor is one factor with 2 conditions
- No blocking factor / corrections

# Example 1: Does dark respiration differ between C3 and C4 plants?

$response = A + D \times predictor + \epsilon$

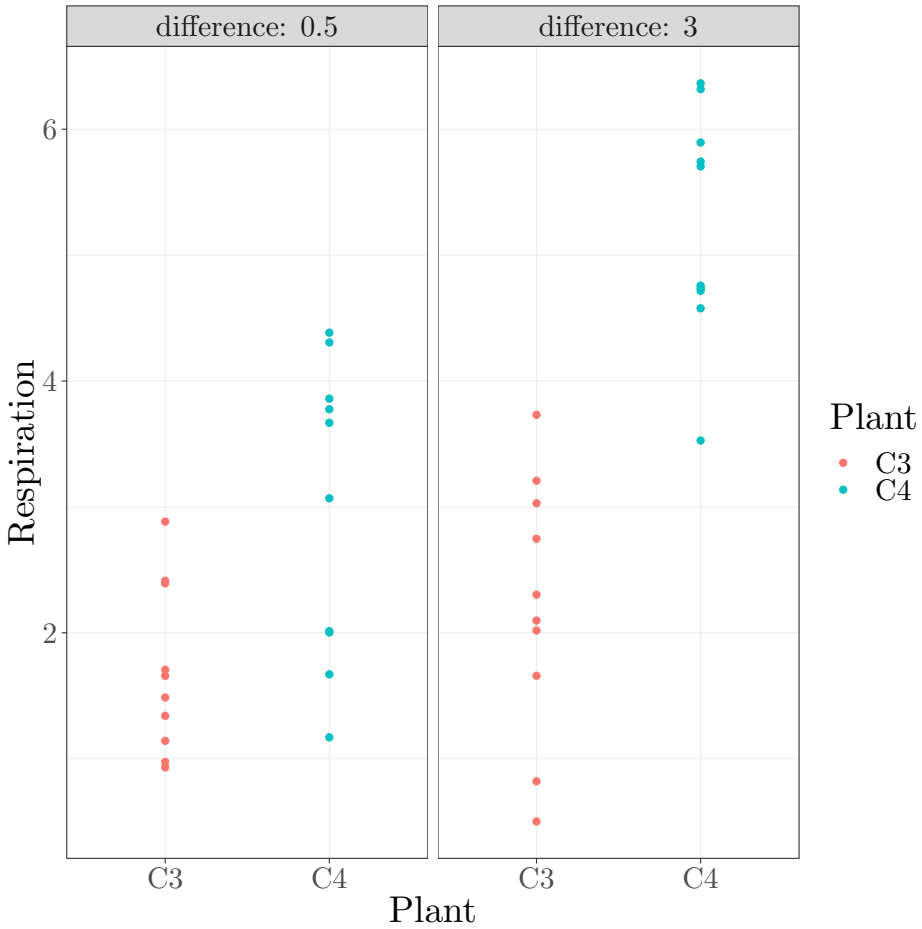
Can we separate the signal $D$ from the noise $\epsilon$ ?

## T-test

- Outcome is a continuous variable
- Experimental factor is one factor with 2 conditions
- No blocking factor / corrections

$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

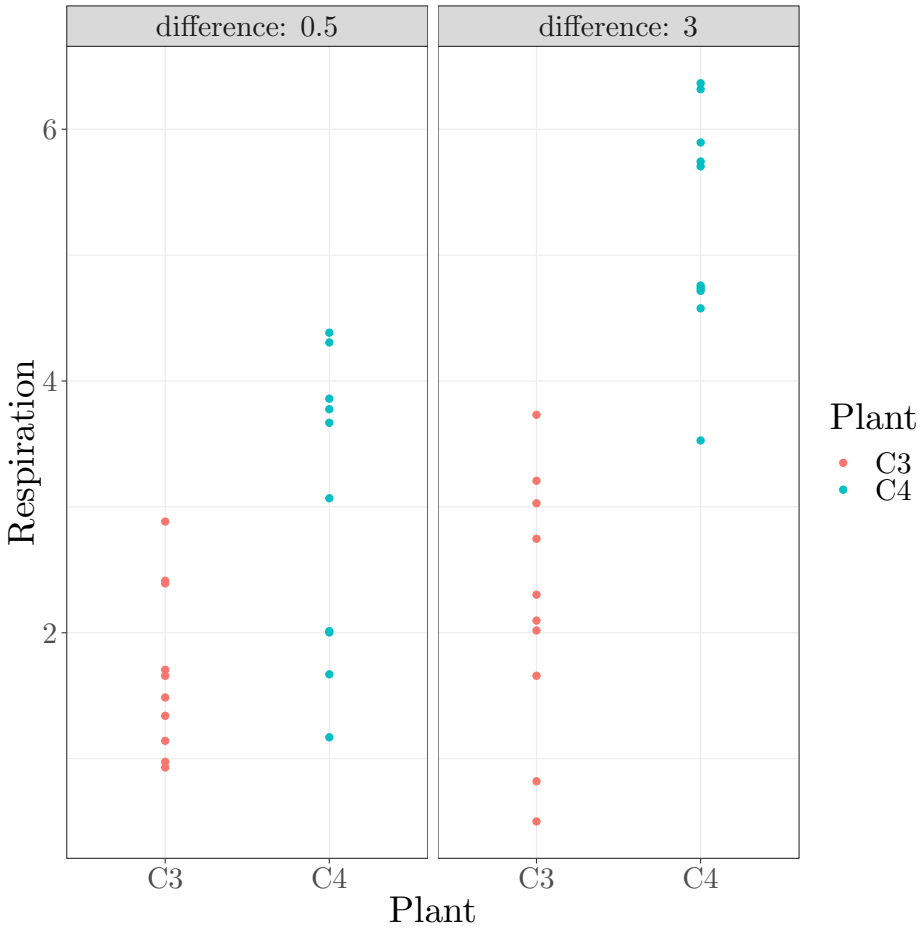
# When can we know whether $D \neq 0$ ?




$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

Is it easier when the true difference is 0.5 or when it is 3 ?

# When can we know whether $D \neq 0$ ?

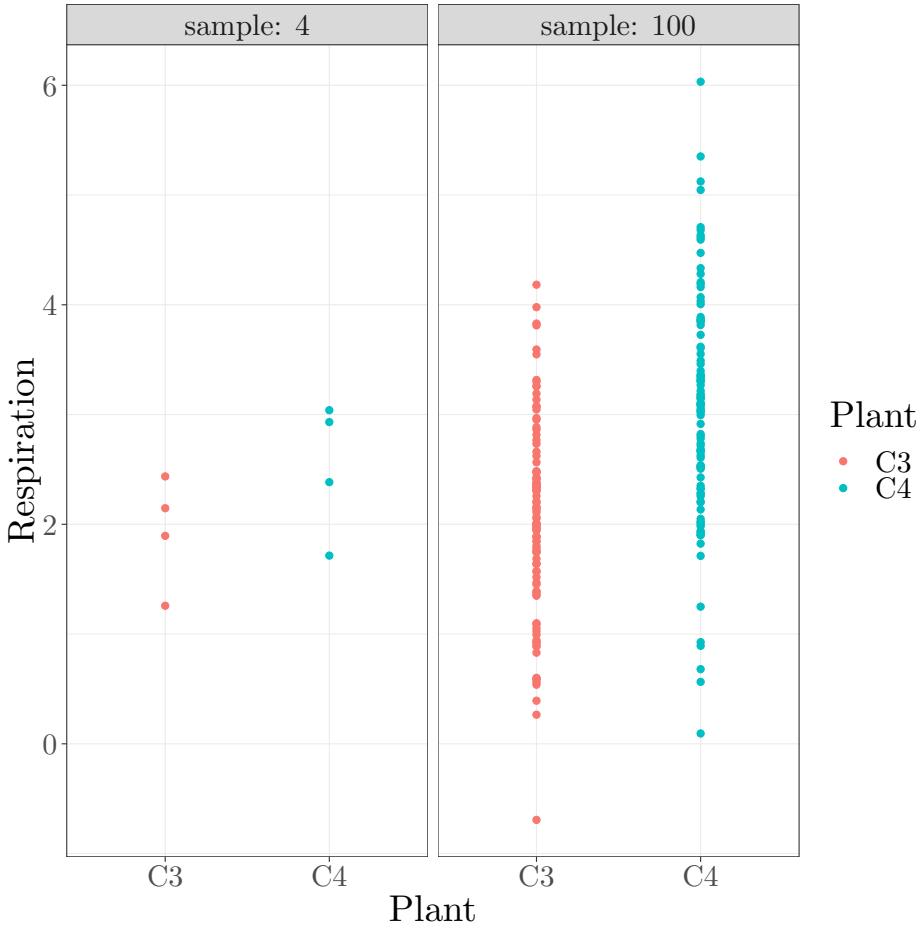



$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

Is it easier when the true difference is 0.5 or when it is 3 ?

1. Large true difference between the means

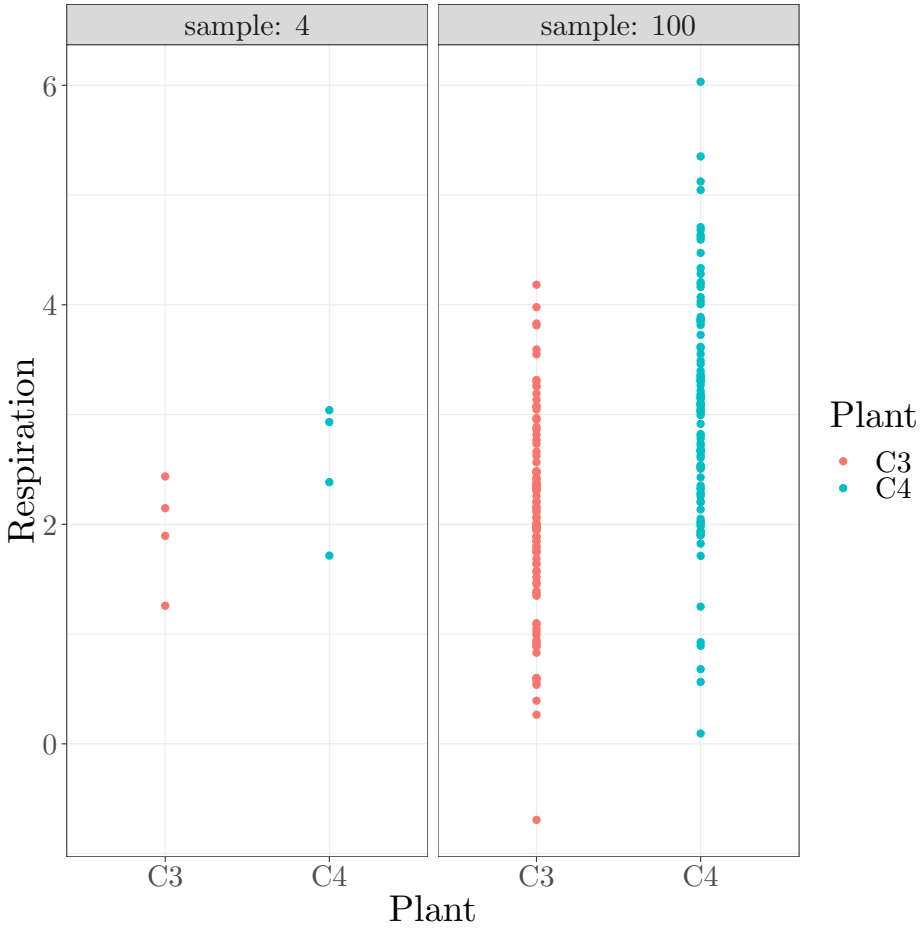
# When can we know whether $D \neq 0$ ?




$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

Is it easier when sample size is 4 or when it is 100?

# When can we know whether $D \neq 0$ ?

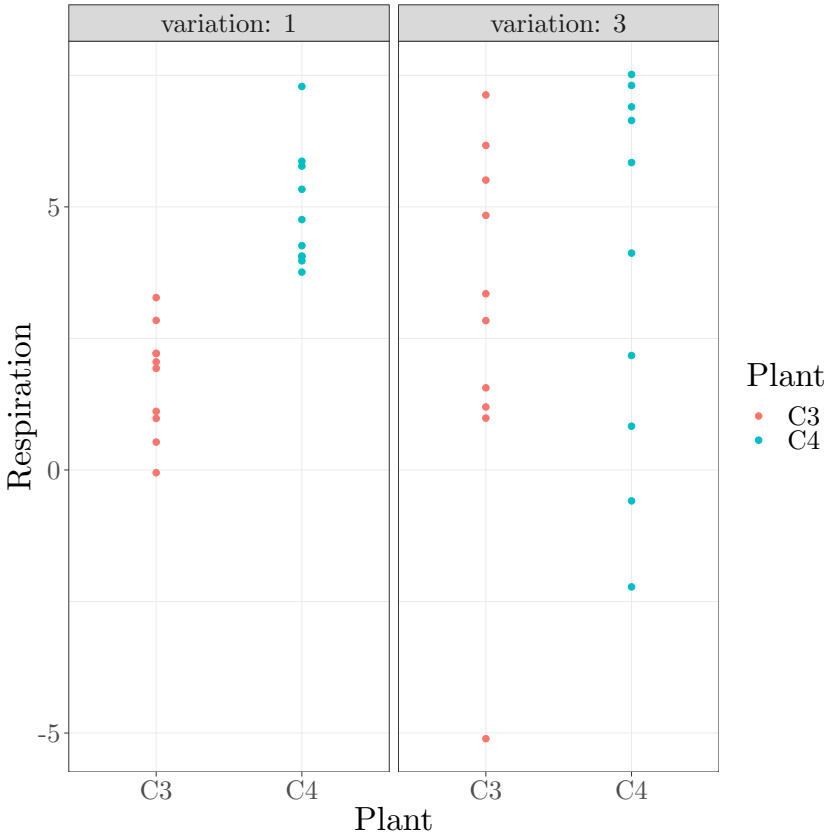



$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

Is it easier when sample size is 4 or when it is 100?

1. Large true difference between the means
2. Large sample size

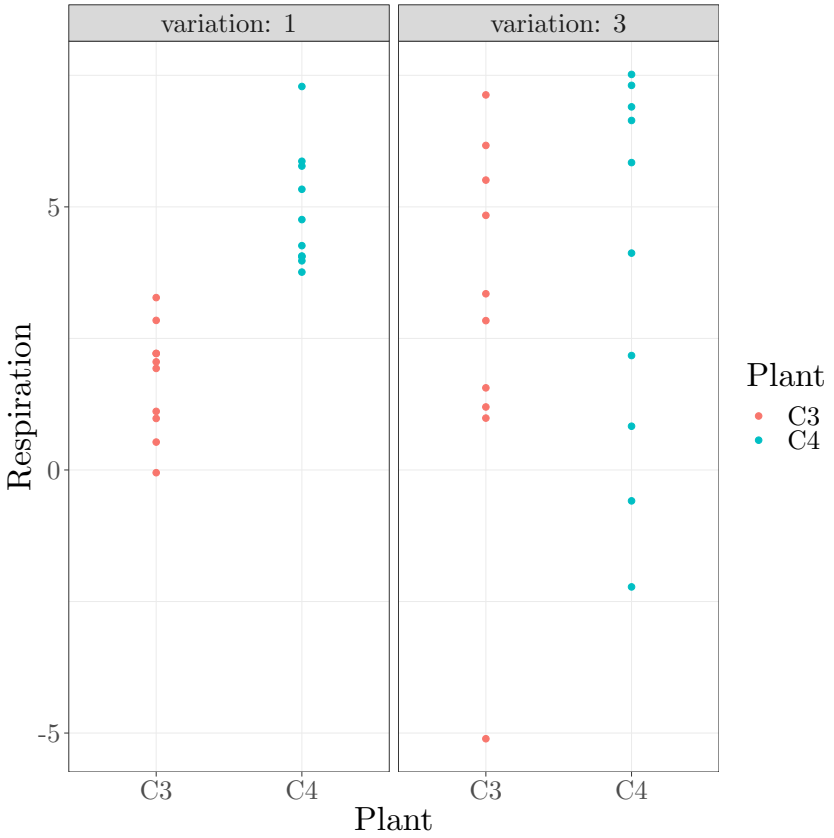
# When can we know whether $D \neq 0$ ?




$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

Is it easier when unexplained variation is 1 or when it is 3?

# When can we know whether $D \neq 0$ ?



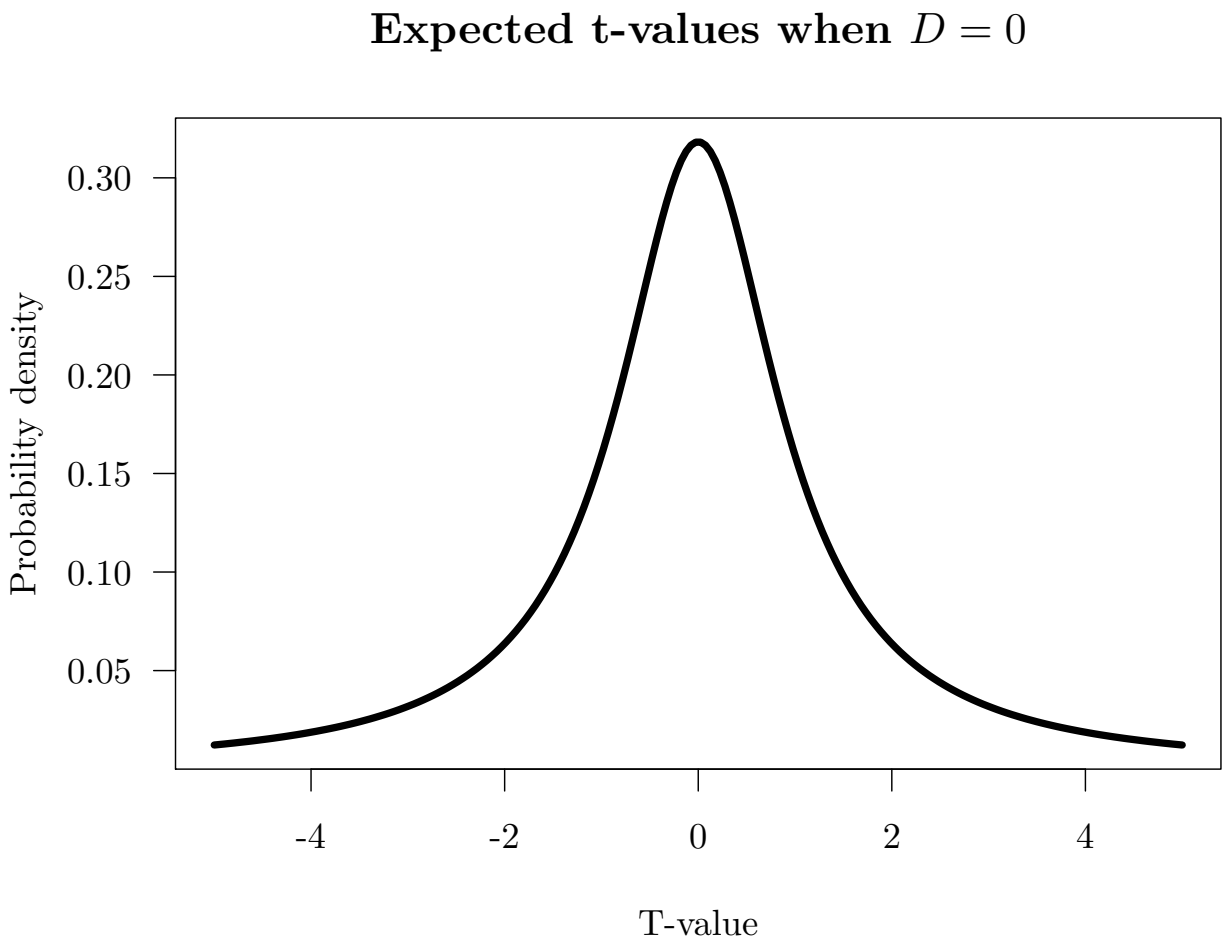

$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

Is it easier when unexplained variation is 1 or when it is 3?

## What makes $t$ large:

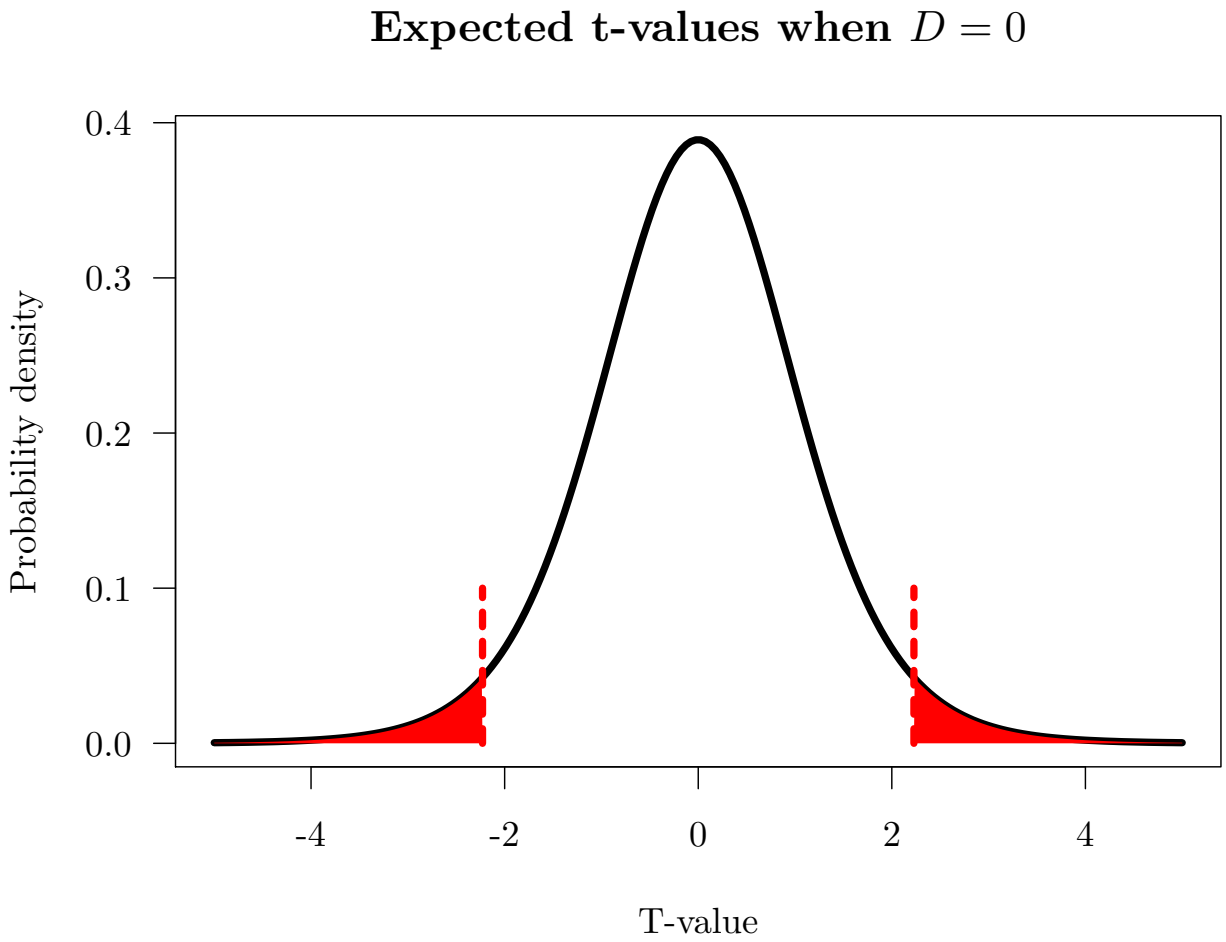
1. Large true difference between the means
2. Large sample size
3. Small unexplained variation

# When can we know whether $D \neq 0$ ?



# When can we know whether $D \neq 0$ ?

**p-value**: probability (area under curve) of getting a value as extreme as what you observed, when D=0



# But really, what is a p-value?

## Candy practical

You got 5 Halloween candies out of the bag. Does the bag contain more Halloween than normal candies?

# Back to C3/C4 plants. Analyse real data in R

1. Set working directory (`setwd(‘‘ / ’’)`) or create a R-project
2. Load and check data

```
resp <- read.csv("d_respiration.csv")
str(resp)
View(resp)
```

3. Visualize data

```
library(ggplot2)
ggplot(resp,aes(Plant_type,rrarea,colour=Plant_type))+
    geom_point()+facet_wrap(~Variation)
```

# Fit a t-test in R: `t.test()`

**Subset data by Variation (High and Low)**

```
resp_H <- subset(resp,Variation == "High")
resp_L <- subset(resp,Variation == "Low")
```

# Fit a t-test in R: `t.test()`

**Subset data by Variation (High and Low)**

```
resp_H <- subset(resp,Variation == "High")
resp_L <- subset(resp,Variation == "Low")
```

**Compare C3 and C4 plants in "High Variation" subset**

```
t.test(rrarea~Plant_type, data=resp_H, var.equal=TRUE)
```

# Fit a t-test in R: `t.test()`

## Subset data by Variation (High and Low)

```
resp_H <- subset(resp,Variation == "High")
resp_L <- subset(resp,Variation == "Low")
```

## Compare C3 and C4 plants in "High Variation" subset

```
t.test(rrarea~Plant_type, data=resp_H, var.equal=TRUE)
```

```
Two Sample t-test
data:  rrarea by Plant_type
t = -0.93776, df = 10, p-value = 0.3705
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.7619349  0.7181446
sample estimates:
mean in group C3 mean in group C4
        2.720021         3.241916
```

# Fit a t-test in R: t.test()

## Subset data by Variation (High and Low)

```
resp_H <- subset(resp,Variation == "High")
resp_L <- subset(resp,Variation == "Low")
```

## Compare C3 and C4 plants in "High Variation" subset

```
t.test(rrarea~Plant_type, data=resp_H, var.equal=TRUE)
```

```
Two Sample t-test
data:  rrarea by Plant_type
t = -0.93776, df = 10, p-value = 0.3705
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.7619349  0.7181446
sample estimates:
mean in group C3 mean in group C4
        2.720021         3.241916
```

# Fit a t-test in R: `t.test()`

**Compare C3 and C4 plants in "Low Variation" subset**

```
t.test(rrarea~Plant_type, data=resp_L, var.equal=TRUE)
```

# Fit an anova in R: aov()

```
aov1 <- aov(rrarea~Plant_type, data=resp_H)
summary(aov1)
```

# Fit an anova in R: aov()

```
aov1 <- aov(rrarea~Plant_type, data=resp_H)
summary(aov1)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Plant_type    1  0.817  0.8171   0.879   0.37
Residuals    10  9.292  0.9292
```

# Fit an anova in R: aov()

```
aov1 <- aov(rrarea~Plant_type, data=resp_H)
summary(aov1)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Plant_type    1  0.817  0.8171   0.879   0.37
Residuals    10  9.292  0.9292
```

$$response = A + D \times predictor + \epsilon$$

# Fit a linear model in R: `lm()`

```
lm1<-lm(rrarea ~ Plant_type, data = resp_L)
summary(lm1)
```

# Fit a linear model in R: `lm()`

```
lm1<-lm(rrarea ~ Plant_type, data = resp_L)
summary(lm1)
```

```
lm(formula = rrarea ~ Plant_type, data = resp_H)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7380 -0.4201 -0.1437  0.6706  1.6754

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.7200     0.3935   6.912 4.13e-05 ***
Plant_typeC4   0.5219     0.5565   0.938     0.37
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9639 on 10 degrees of freedom
Multiple R-squared:  0.08083,Adjusted R-squared:  -0.01109
F-statistic: 0.8794 on 1 and 10 DF,  p-value: 0.3705
```

# Fit a linear model in R: `lm()`

```
lm1<-lm(rrarea ~ Plant_type, data = resp_L)
summary(lm1)
```

```
lm(formula = rrarea ~ Plant_type, data = resp_H)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7380 -0.4201 -0.1437  0.6706  1.6754

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.7200     0.3935   6.912 4.13e-05 ***
Plant_typeC4   0.5219     0.5565   0.938     0.37
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9639 on 10 degrees of freedom
Multiple R-squared:  0.08083,Adjusted R-squared:  -0.01109
F-statistic: 0.8794 on 1 and 10 DF,  p-value: 0.3705
```

# Fit a linear model in R: `lm()`

```
library(emmeans)
emmeans(lm1, ~Plant_type)
```

```
 Plant_type   emmean        SE df lower.CL upper.CL
 C3         2.720021 0.3935305 10 1.843180 3.596861
 C4         3.241916 0.3935305 10 2.365076 4.118757

Confidence level used: 0.95
```

# Fit a linear model in R: `lm()`

```
library(emmeans)
emmeans(lm1, ~Plant_type)
```

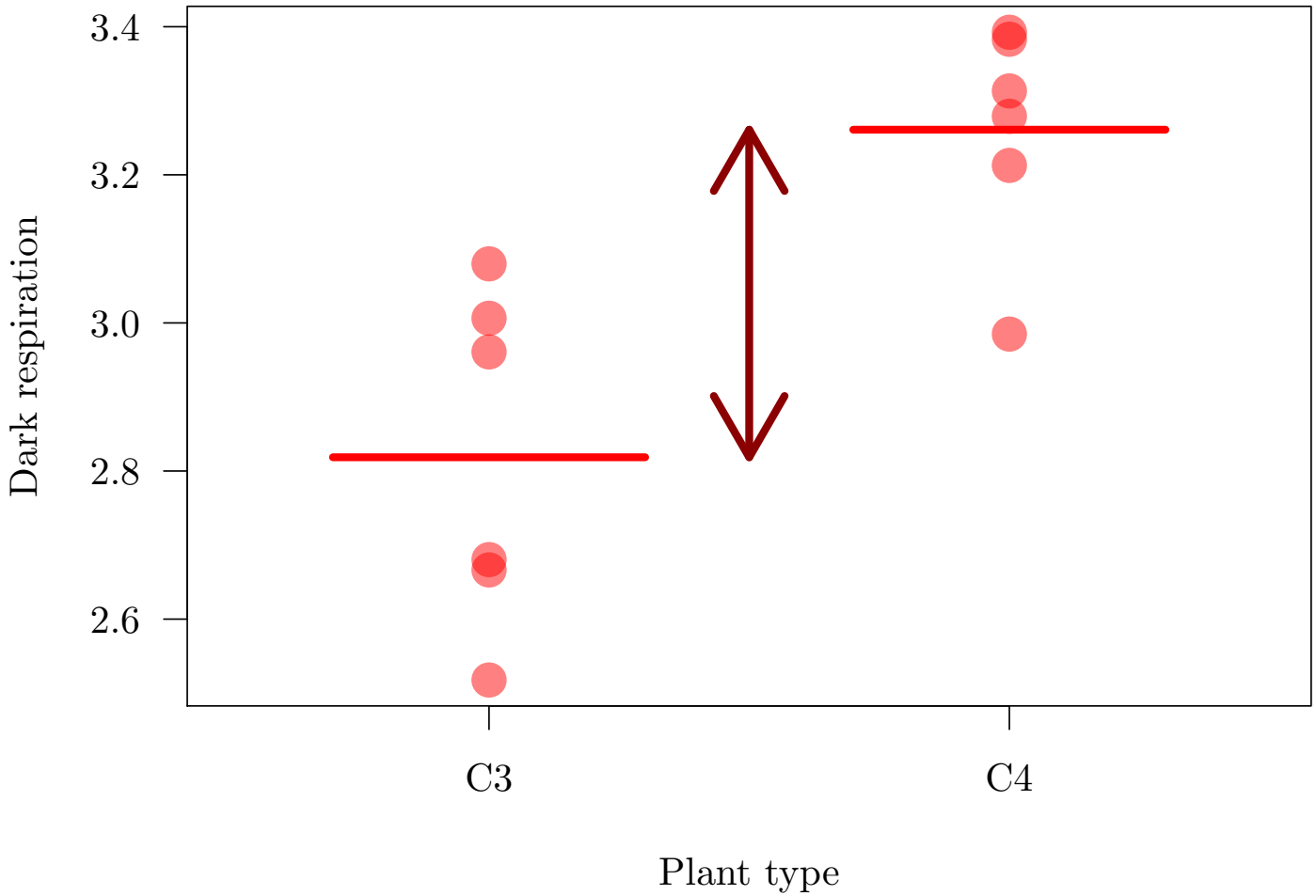
```
 Plant_type   emmean        SE df lower.CL upper.CL
 C3          2.720021 0.3935305 10 1.843180 3.596861
 C4          3.241916 0.3935305 10 2.365076 4.118757

Confidence level used: 0.95
```

$$response = A + D \times predictor + \epsilon$$

# Compare the output from t.test, aov and lm

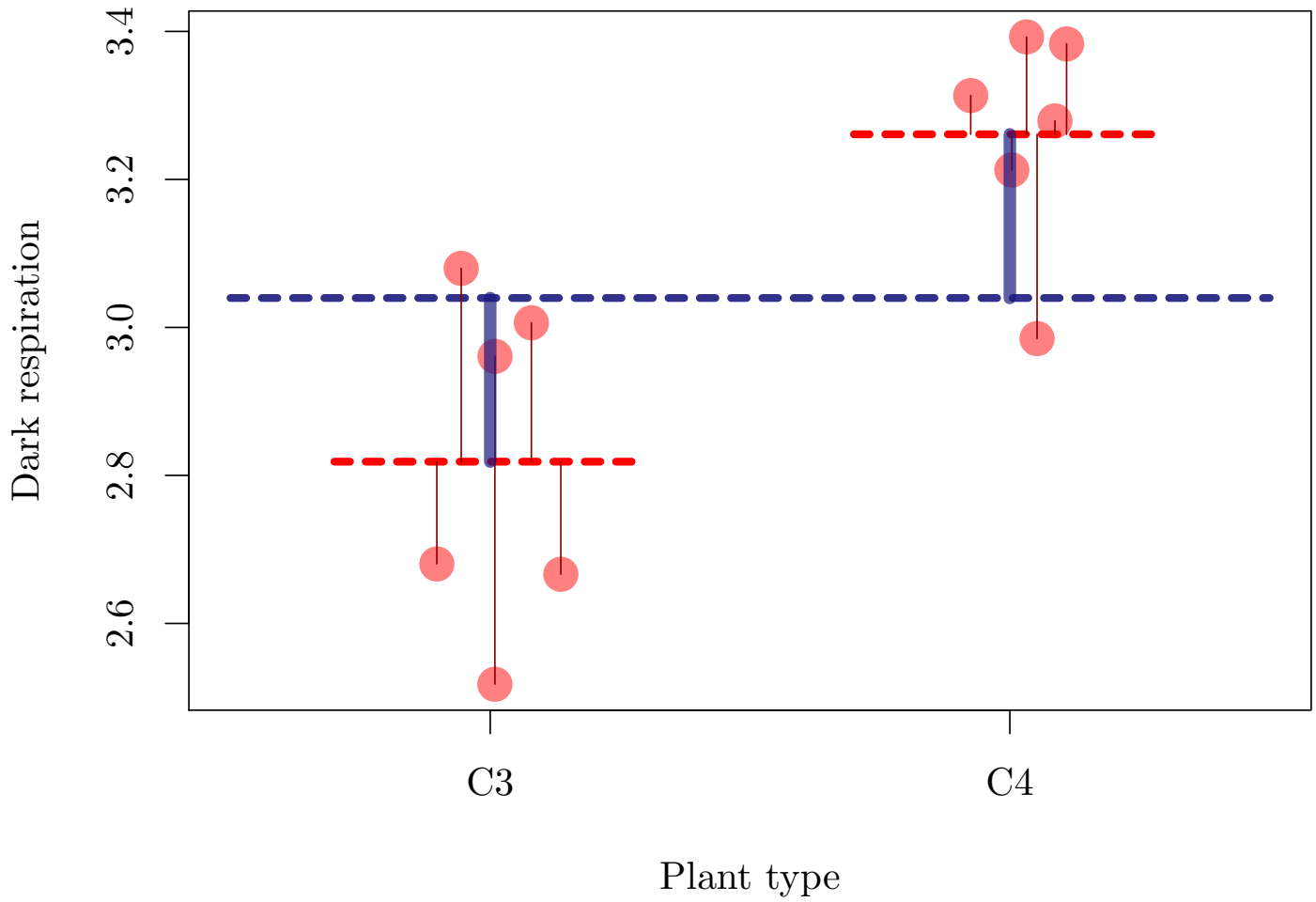
# Three equivalent ways to look at data

T-test, focus on difference between two means



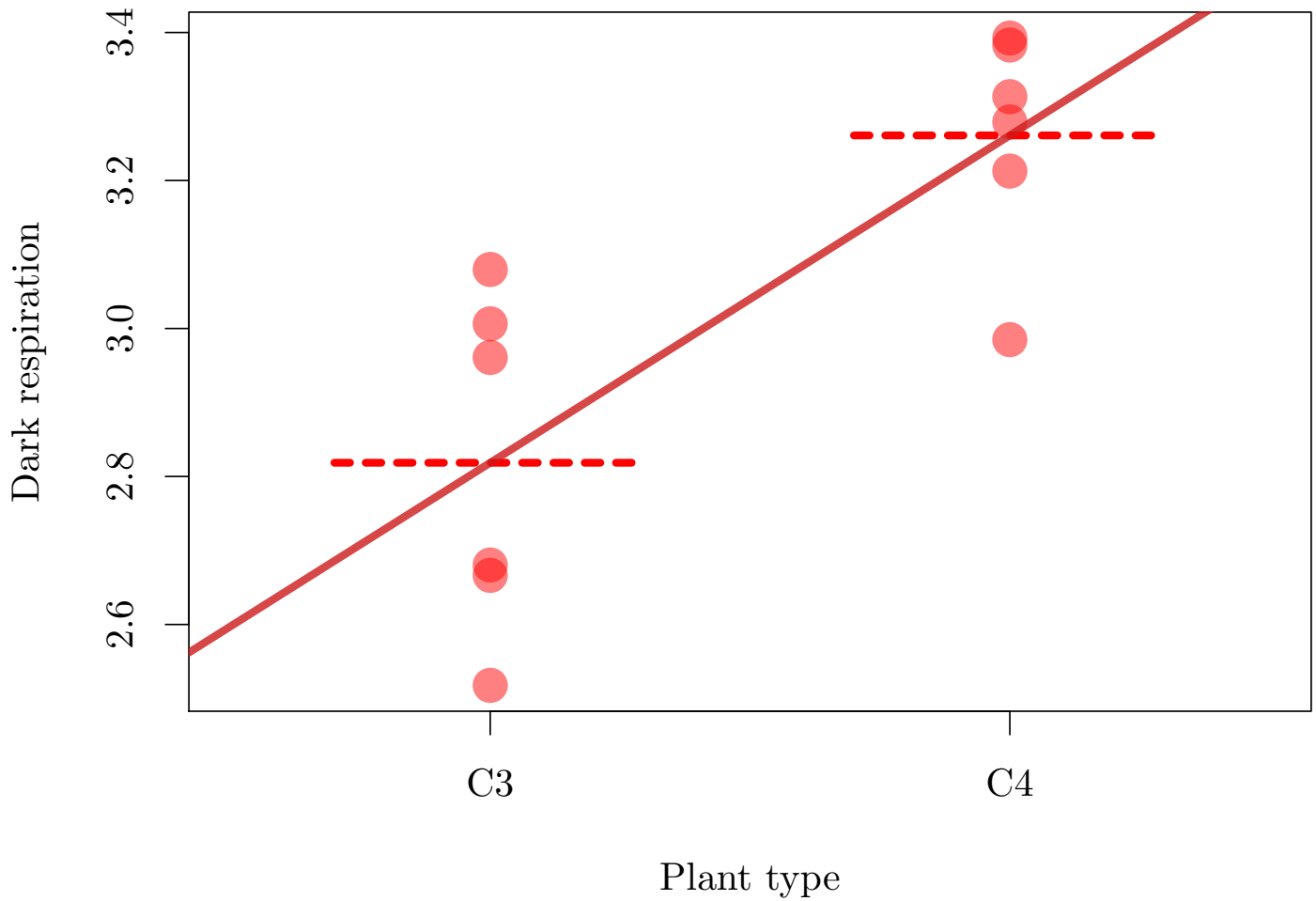
# Three equivalent ways to look at data

ANOVA, focus on variation within VS. between



# Three equivalent ways to look at data

Linear regression, focus on rate of change



# All is one. . .

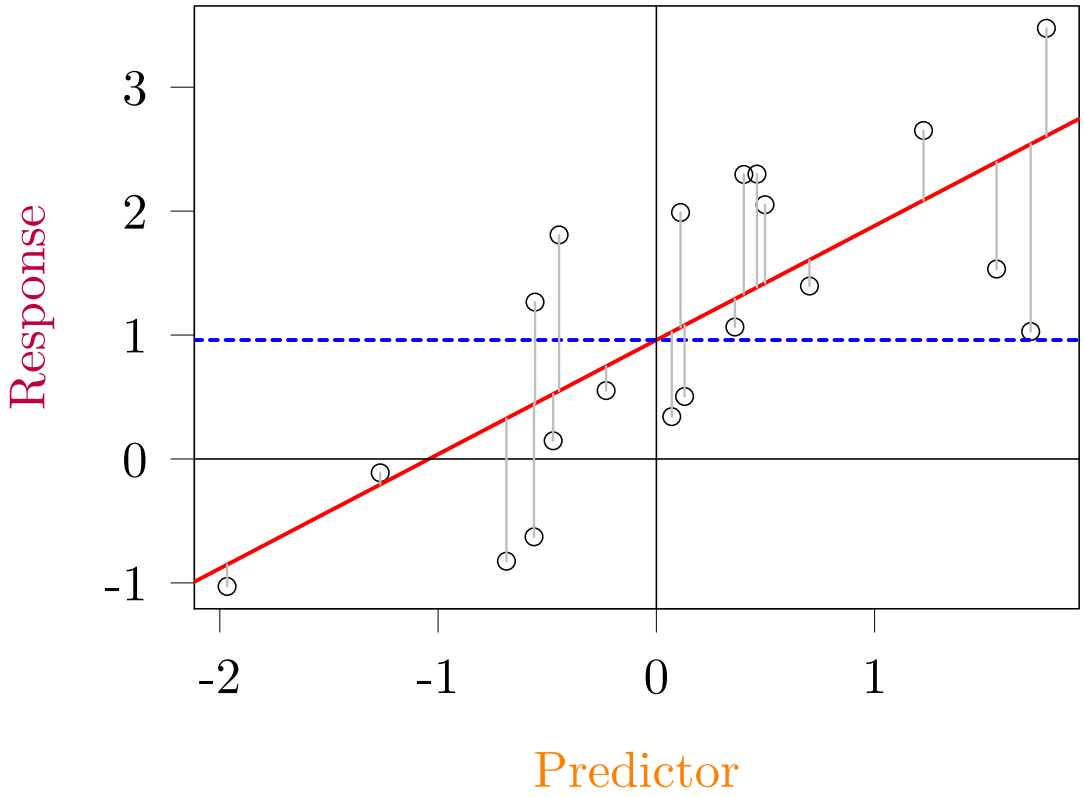
# All is one...

## ...but `lm()` rules (IMH)

- t-test, ANOVA, regression and others can be mathematically equivalent
- In R, `lm()` and related functions can do them all...
- ...and much more!

# Focus on linear models

**Response** = **Intercept** + **Slope** × **Predictor** + **Error**



# A simple linear model

**Response** = **Intercept** + **Slope** × **Predictor** + **Error**

```
lm(response ~ 1 + predictor1 + predictor2, data=data)
```

equivalent to

```
lm(response ~ predictor1 + predictor2, data=data)
```

equivalent to

```
lm(response ~ predictor2 + predictor1, data=data)
```

- Intercept can be explicit or implicit
- Can remove intercept with `...~ 0 + ...`
- Error is implicit
- Feed the option `data=` to keep code short, reliable and flexible
- Order of predictors do not matter

1. Cautionary tales from the front
2. Introduction to Statistical Modelling
3. Another look at essential steps

# General approach

1. Scientific question

# General approach

1. Scientific question
↓
2. Model and Statistical question

# General approach

1. Scientific question

$\downarrow$

2. Model and Statistical question $\longleftrightarrow$ 3. Data collection

# General approach

1. Scientific question

↓

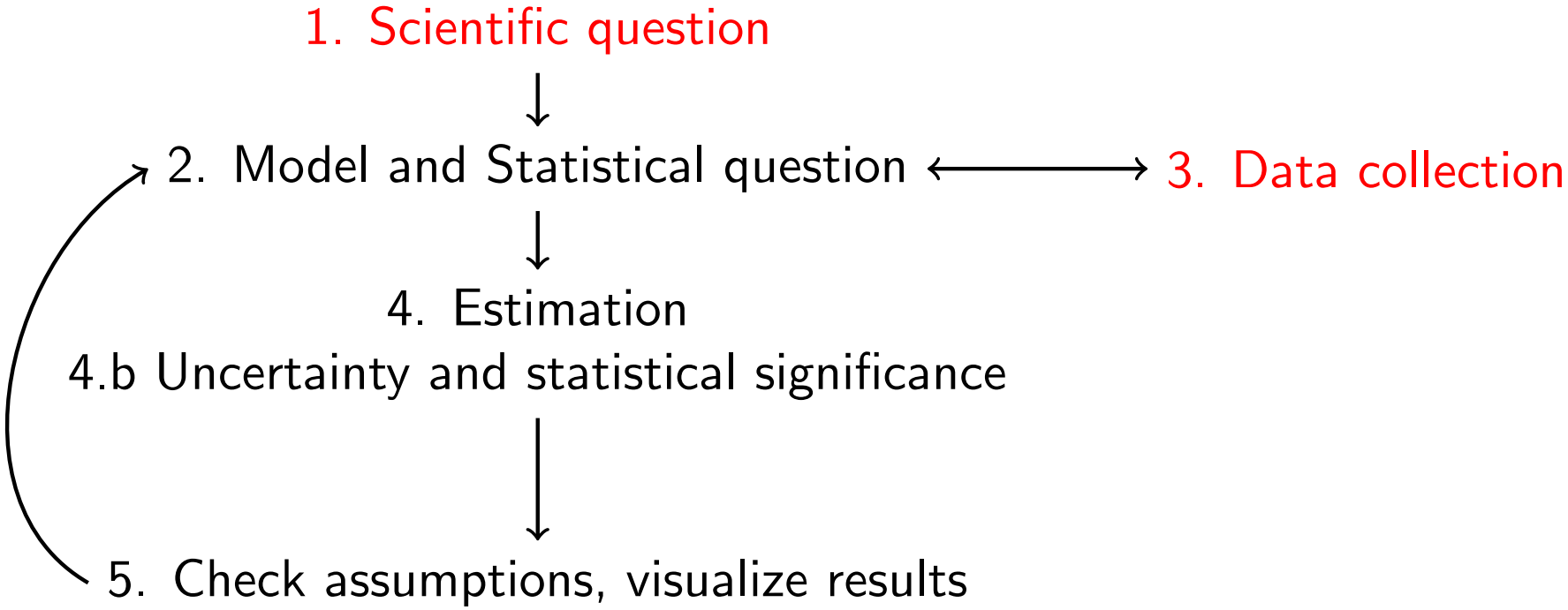
2. Model and Statistical question ⟷ 3. Data collection

↓

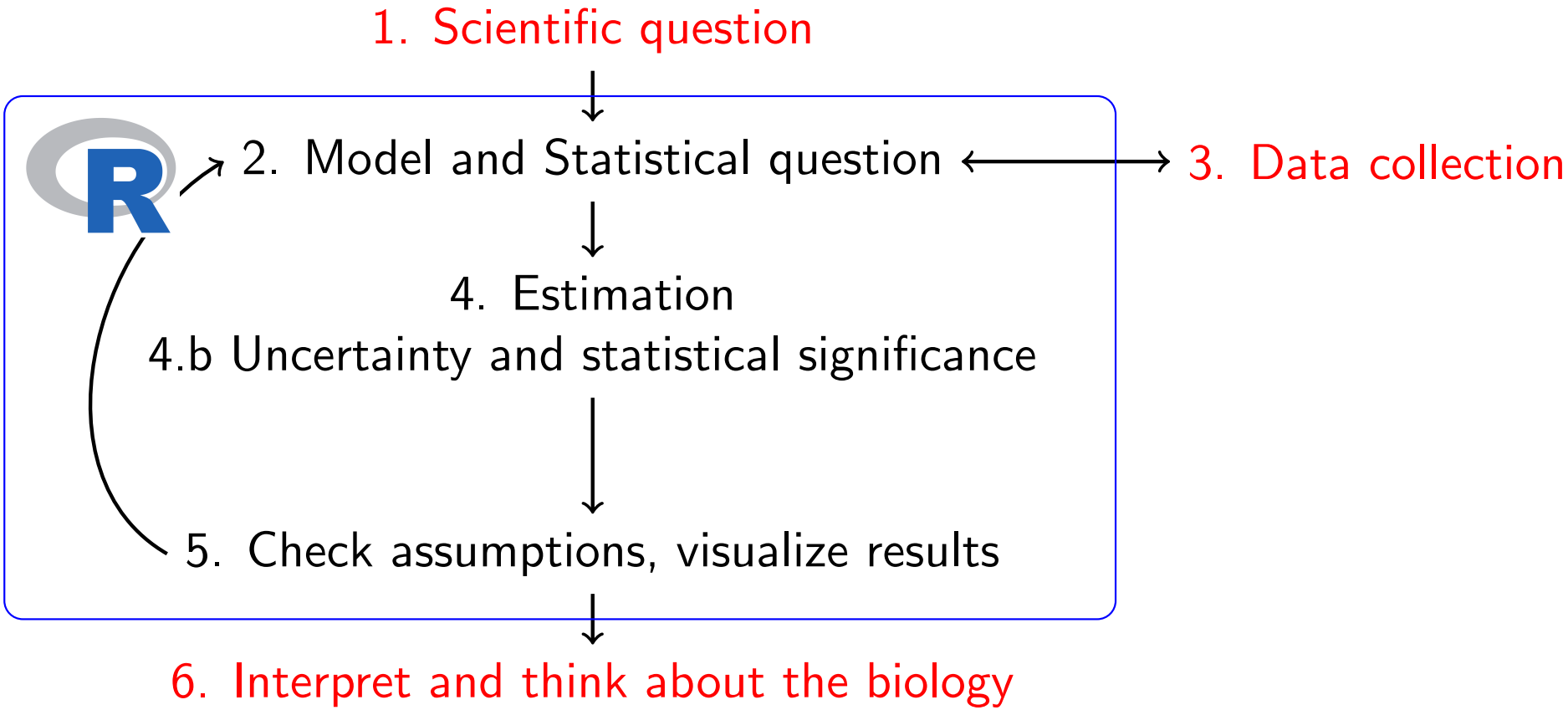
4. Estimation

4.b Uncertainty and statistical significance

# General approach



# General approach



# Back to C3/C4

```
lmL<-lm(rrarea ~ Plant_type, data = resp_L)
summary(lmL)
```

# Back to C3/C4

```
lmL<-lm(rrarea ~ Plant_type, data = resp_L)
summary(lmL)
```

```
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.81857    0.07856  35.878 6.72e-12 ***
Plant_typeC4  0.44235    0.11110   3.982  0.00259 ** ---
...
```

**Estimation:**

$response = A + D \times predictor + \epsilon$

$A$ =?, $D$ =?

# Back to C3/C4

```
lmL<-lm(rrarea ~ Plant_type, data = resp_L)
summary(lmL)
```

```
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.81857    0.07856  35.878 6.72e-12 ***
Plant_typeC4  0.44235    0.11110   3.982  0.00259 ** ---
...
```

**Estimation:**

$response = A + D \times predictor + \epsilon$

$A = 2.81857$, $D = 0.44235$

# Back to C3/C4

```
lmL<-lm(rrarea ~ Plant_type, data = resp_L)
summary(lmL)
```

```
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.81857    0.07856  35.878 6.72e-12 ***
Plant_typeC4  0.44235    0.11110   3.982  0.00259 ** ---
...
```

**Estimation:**

$response = A + D \times predictor + \epsilon$

$A = 2.81857$, $D = 0.44235$

**Uncertainty:**

For $D$ SE= 0.11110 ; p-value=0.00259

# Back to C3/C4

```
lmL<-lm(rrarea ~ Plant_type, data = resp_L)
summary(lmL)
```

```
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.81857    0.07856  35.878 6.72e-12 ***
Plant_typeC4  0.44235    0.11110   3.982  0.00259 ** ---
...
```

**Estimation:**

$response = A + D \times predictor + \epsilon$

$A = 2.81857$, $D = 0.44235$

**Uncertainty:**

For $D$ SE= 0.11110 ; p-value=0.00259

**What do we do next?**

# Check assumptions, visualize results

# Check assumptions, visualize results

## Linear model basic assumptions

- Predictor not perfectly correlated
  *Risk: Model won't run, unstable convergence, or huge SE*

# Check assumptions, visualize results

## Linear model basic assumptions

- Predictor not perfectly correlated
  *Risk: Model won't run, unstable convergence, or huge SE*
- Little error in predictors
  *Risk: bias estimates (underestimate with Gaussian error)*

# Check assumptions, visualize results

## Linear model basic assumptions

- Predictor not perfectly correlated
  *Risk: Model won't run, unstable convergence, or huge SE*
- Little error in predictors
  *Risk: bias estimates (underestimate with Gaussian error)*
- Gaussian error distribution
  *Risk: Poor predictions*

# Check assumptions, visualize results

## Linear model basic assumptions

- Predictor not perfectly correlated
  *Risk: Model won't run, unstable convergence, or huge SE*
- Little error in predictors
  *Risk: bias estimates (underestimate with Gaussian error)*
- Gaussian error distribution
  *Risk: Poor predictions*
- Homoscedasticity (constant error variance)
  *Risk: Over-optimistic uncertainty, unreliable predictions*

# Check assumptions, visualize results

## Linear model basic assumptions

- Predictor not perfectly correlated
  *Risk: Model won't run, unstable convergence, or huge SE*
- Little error in predictors
  *Risk: bias estimates (underestimate with Gaussian error)*
- Gaussian error distribution
  *Risk: Poor predictions*
- Homoscedasticity (constant error variance)
  *Risk: Over-optimistic uncertainty, unreliable predictions*
- Independence of error
  *Risk: Bias and over-optimistic uncertainty*

# Check assumptions, visualize results

Assessing model assumptions in R:

```
lmL<-lm(rrarea ~ Plant_type, data = resp_L)
plot(lmL)
summary(lmL)
```

# Check assumptions, visualize results

## Visualize and report results

```
lm1.results<-summary(emmeans(lm1,~Plant_type))

ggplot(lm1.results,aes(Plant_type,emmean, fill=Plant_type))+
  geom_bar(stat="identity", width=.4)+
  geom_errorbar(aes(ymin =lm1.results$lower.CL,
  ymax = lm1.results$upper.CL), width=.2)+
  ylim(0,4)+
  geom_point(data=resp_L, aes(x=Plant_type, y=rrarea), color="red")+
  labs(y = "Dark Respiration (units)")+
  geom_text(aes(x=1.5, y=3.5, label="p=0.002"))
```

# Check assumptions, visualize results