

Statistical Thinking in Biology Research

Chapter 1

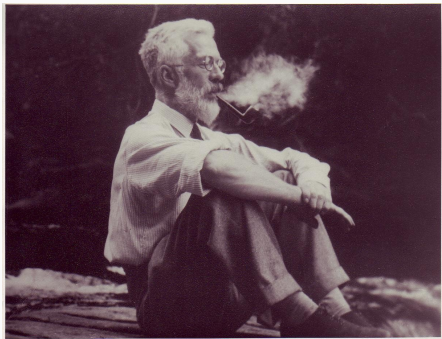
Timothée Bonnet & Terry Neeman

May 13, 2019

Research School Biology and Biological Data Science Institute

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)

The big picture

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

Sir Ronald Fisher

Presidential Address to the First Indian Statistical Congress, 1938. Sankhya 4, 14-17

The big picture

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

Sir Ronald Fisher

Presidential Address to the First Indian Statistical Congress, 1938. Sankhya 4, 14-17

You won't need to call in a statistician too often if you are (almost) one yourself

The big picture

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

Sir Ronald Fisher

Presidential Address to the First Indian Statistical Congress, 1938. Sankhya 4, 14-17

You won't need to call in a statistician too often if you are (almost) one yourself

Statistics is...

1. interesting
2. a unifying language of sciences
3. empowering

Approximate plan

Monday morning Cautionary tales, General approach to modelling

Monday afternoon Experimental design

Tuesday morning Mean structure, *additive effects and interactions*

Tuesday afternoon Variance structure, *mixed models*

Wednesday Data generating process, *GLMs*; + practice with your data?

If you want, send me:

- Your data / your planned experiment
- Brief explanation of biological system and question

... we will look at it together

Key ideas for today

- Statistics in biology = study biological variation

Key ideas for today

- Statistics in biology = study biological variation
- Understanding statistical ideas about biological variation:

Key ideas for today

- Statistics in biology = study biological variation
- Understanding statistical ideas about biological variation:
 - Informs the design of experiments

Key ideas for today

- Statistics in biology = study biological variation
- Understanding statistical ideas about biological variation:
 - Informs the design of experiments
 - Informs the analysis of experiments

Key ideas for today

- Statistics in biology = study biological variation
- Understanding statistical ideas about biological variation:
 - Informs the design of experiments
 - Informs the analysis of experiments
- Statistical thinking is an essential component of scientific thinking

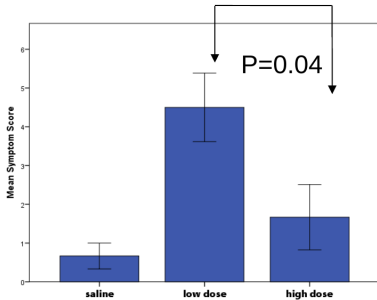
Cautionary tales from the front

Introduction to Statistical Modelling

Another look at essential steps

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

Mean symptom score by treatment group



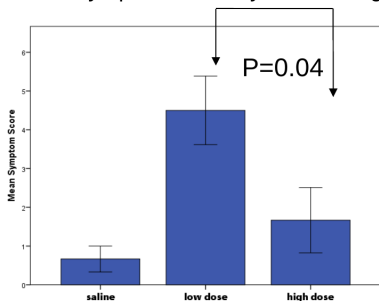
Vaccine challenge experiment:

- 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

Mean symptom score by treatment group



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



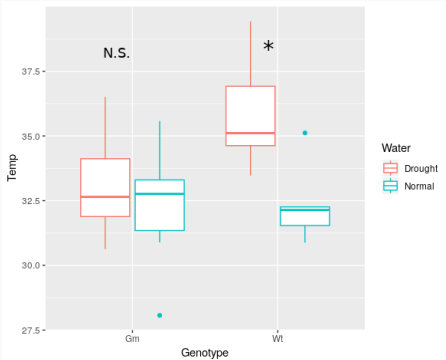
Cage 2:
Low Dose



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”

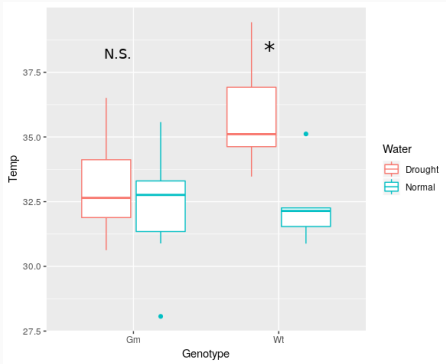


Are temperature mechanisms modified in a genetically modified tomato plant?

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

T-tests between water conditions:
GM: $p=0.46$; Wt: $p=0.02$

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

T-tests between water conditions:

GM: $p=0.46$; Wt: $p=0.02$

Evidence of difference + No evidence of difference

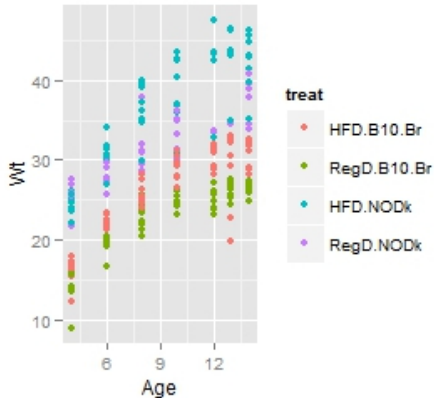
\neq

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)

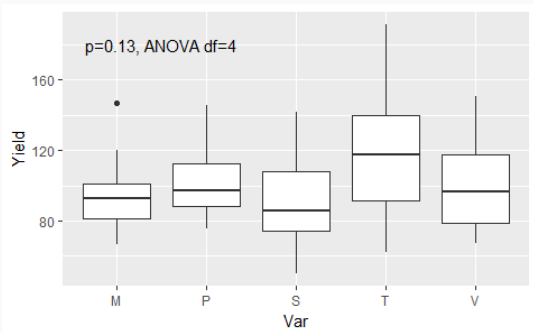
Acknowledgements: Ainy Hussain, PhD student 2013

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

Message 4: Knowing how to combine information across sub-groups can improve inference

Comparing yield in five barley varieties (1930s)

Experimental factors: 5 varieties of barley, 6 locations, 2 time points. Outcome measure: yield

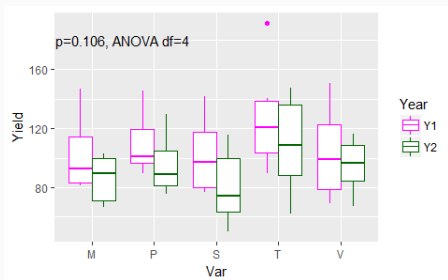


Acknowledgements: MASS R-package

Message 4: Knowing how to combine information across sub-groups 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:

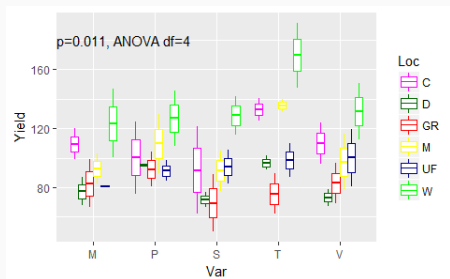
- Controlling for year = comparing yield WITHIN years and combining these

Acknowledgements: MASS R-package

Message 4: Knowing how to combine information across sub-groups can improve inference

Comparing yield in five barley varieties (1930s)

Experimental factors: 5 varieties of barley, 6 locations, 2 time points. Outcome measure: yield

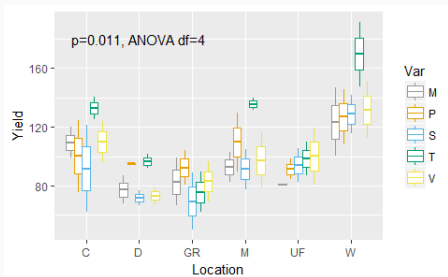


Controlling for other sources of variation:

- 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 sub-groups 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 with different temperatures
- Response - % germinated

What factors other than temperature to consider?

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.**

Cautionary tales from the front

Introduction to Statistical Modelling

Another look at essential steps

Introduction to Statistical Modelling

What is a statistical model?

What is a statistical model?

- A formal but simplified representation of the world / an experiment
- If the representation is good enough for our goal...
- ... we can learn something from data

Introduction to Statistical Modelling

What is a statistical model?

- A formal but simplified representation of the world / an experiment
- If the representation is good enough for our goal...
- ... we can learn something from data



Key components of a statistical model of an experiment

1. Outcome measure

- Response variable
- Measure of interest

2. Experimental factors

- Conditions that can be manipulated
- Conditions of interest (e.g. genotype, gender)
- Main questions: do the conditions impact upon the outcome measure?

3. Blocking factors

- Conditions (not of interest) that may impact upon the outcome
- Sources of variation in the experiment that need to be controlled for
- Clustering of experimental units

Key components of a statistical model of an experiment

1. Outcome measure

- Response variable
- Measure of interest

2. Experimental factors

- Conditions that can be manipulated
- Conditions of interest (e.g. genotype, gender)
- Main questions: do the conditions impact upon the outcome measure?

3. Blocking factors

- Conditions (not of interest) that may impact upon the outcome
- Sources of variation in the experiment that need to be controlled for
- Clustering of experimental units

ALWAYS BEGIN WITH A RESEARCH QUESTION

=> what is outcome / exp factors / blocking factors;

=> Do not make-up hypotheses a-posteriori

Key Objectives of a statistical model of an experiment

- Compare mean response across 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?

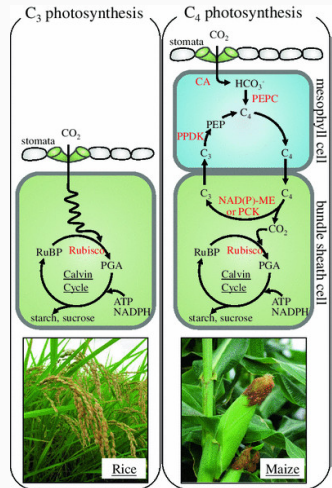
Outcome measure: dark respiration

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?

$$\text{response} = A + D \times \text{predictor} + \epsilon$$

Can we separate the signal D from the noise ϵ ?

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

$$\text{response} = A + D \times \text{predictor} + \epsilon$$

Can we separate the signal D from the noise ϵ ?

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?

$$\text{response} = A + D \times \text{predictor} + \epsilon$$

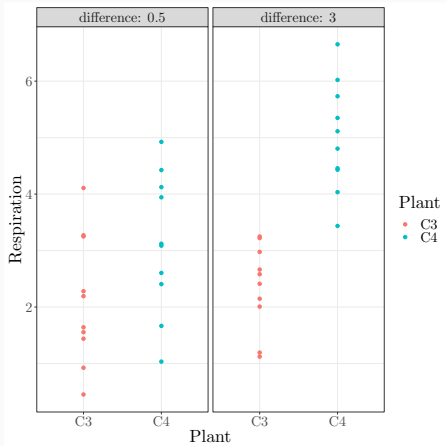
Can we separate the signal D from the noise ϵ ?

T-test

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

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

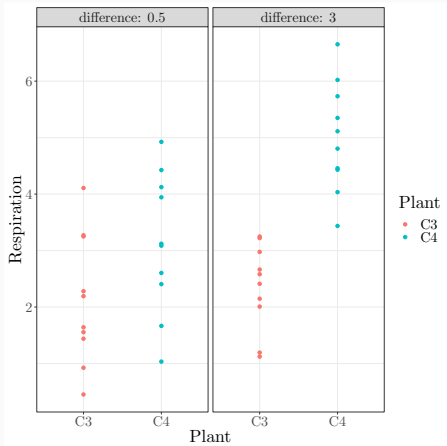
When can we know whether $D \neq 0$?



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

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

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

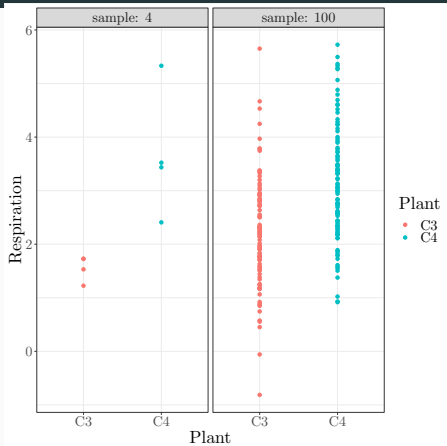


$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

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

1. Large true difference between the means

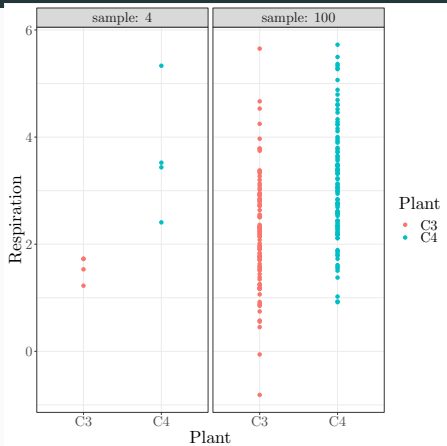
When can we know whether $D \neq 0$?



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

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

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

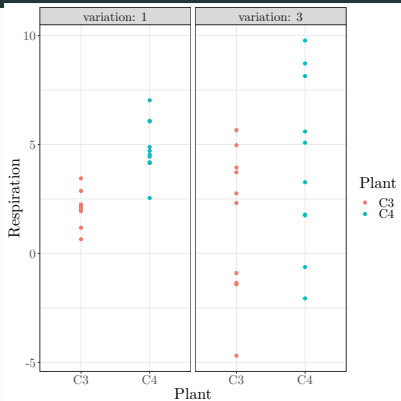


$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

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

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

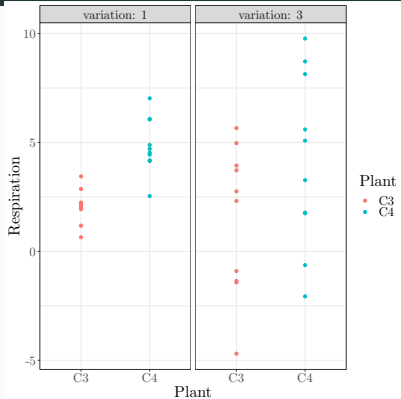
When can we know whether $D \neq 0$?



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

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

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



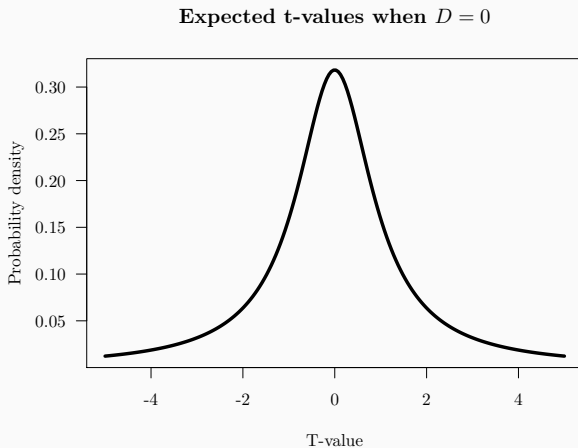
$$t = \frac{D}{\text{Variation of } \epsilon} \times \frac{\text{Sample Size}}{\sqrt{2}}$$

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

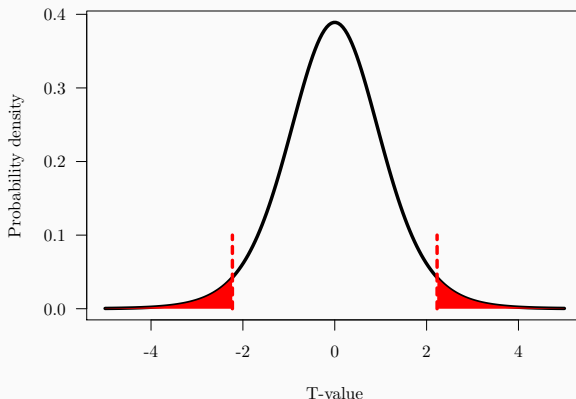
When can we know whether $D \neq 0$?



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 the true $D=0$

Expected t-values when $D = 0$



But really, what is a p-value?

Candy practical

- You got a pack of 20 candies with a mix of Halloween and Fruit candies

But really, what is a p-value?

Candy practical

- You got a pack of 20 candies with a mix of Halloween and Fruit candies
- You pick one, it's a Halloween one. . . looks quite disgusting. You put in back

But really, what is a p-value?

Candy practical

- You got a pack of 20 candies with a mix of Halloween and Fruit candies
- You pick one, it's a Halloween one. . . looks quite disgusting. You put it back
- You pick a second one. Again a disgusting Halloween candy! You put it back

But really, what is a p-value?

Candy practical

- You got a pack of 20 candies with a mix of Halloween and Fruit candies
- You pick one, it's a Halloween one. . . looks quite disgusting. You put it back
- You pick a second one. Again a disgusting Halloween candy! You put it back
- And so on, until 5 candies. You wonder if you have been cheated.

But really, what is a p-value?

Candy practical

- You got a pack of 20 candies with a mix of Halloween and Fruit candies
- You pick one, it's a Halloween one. . . looks quite disgusting. You put it back
- You pick a second one. Again a disgusting Halloween candy! You put it back
- And so on, until 5 candies. You wonder if you have been cheated.
- Are there more Halloween than Fruit candies in that pack?

But really, what is a p-value?

Candy practical

- You got a pack of 20 candies with a mix of Halloween and Fruit candies
- You pick one, it's a Halloween one. . . looks quite disgusting. You put it back
- You pick a second one. Again a disgusting Halloween candy! You put it back
- And so on, until 5 candies. You wonder if you have been cheated.
- Are there more Halloween than Fruit candies in that pack?
- You decide to use statistics to find out

How to?

- Draw 5 candies out of the pack
- Write down how many Halloween candies
- How often is it 5?

But really, what is a p-value?

https://docs.google.com/spreadsheets/d/1Y9512z1xxkphjAZ_dYT9SqfQH02UqTDK1W2X0mDEwZY/edit?usp=sharing

- Draw 5 candies out of the pack
- Write down how many Halloween candies
- How often is it 5?

But really, what is a p-value?

https://docs.google.com/spreadsheets/d/1Y9512z1xxkphjAZ_dYT9SqfQH02UqTDK1W2X0mDEwZY/edit?usp=sharing

- Draw 5 candies out of the pack
- Write down how many Halloween candies
- How often is it 5?
- Estimate the p-value for the test “candies have same frequency”

But really, what is a p-value?

https://docs.google.com/spreadsheets/d/1Y9512z1xxkphjAZ_dYT9SqfQH02UqTDK1W2X0mDEwZY/edit?usp=sharing

- Draw 5 candies out of the pack
- Write down how many Halloween candies
- How often is it 5?
- Estimate the p-value for the test “candies have same frequency”
- Redo the experiment in R, using random sampling (`rbinom`)

But really, what is a p-value?

https://docs.google.com/spreadsheets/d/1Y9512z1xxkphjAZ_dYT9SqfQH02UqTDK1W2X0mDEwZY/edit?usp=sharing

- Draw 5 candies out of the pack
- Write down how many Halloween candies
- How often is it 5?
- Estimate the p-value for the test “candies have same frequency”
- Redo the experiment in R, using random sampling (`rbinom`)
- What is the correct null-distribution?

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		

$$\text{response} = A + D \times \text{predictor} + \epsilon$$

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

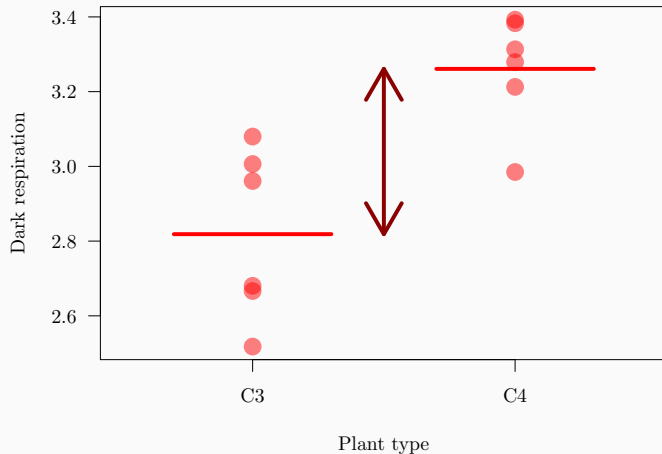
Confidence level used: 0.95

$$\text{response} = A + D \times \text{predictor} + \epsilon$$

Compare the output from t.test, aov and lm

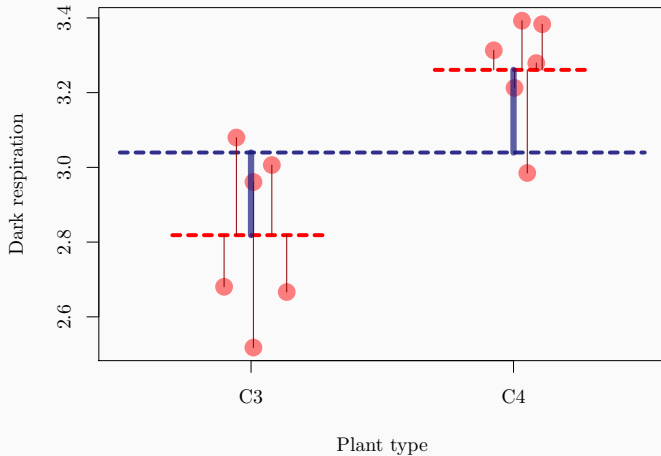
Three equivalent ways to look at data

T-test, focus on difference between two means



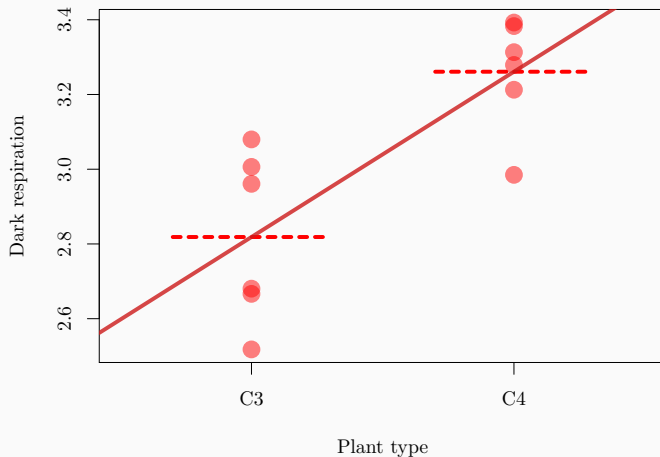
Three equivalent ways to look at data

ANOVA, focus on variation within VS. between



Three equivalent ways to look at data

Linear regression, focus on rate of change

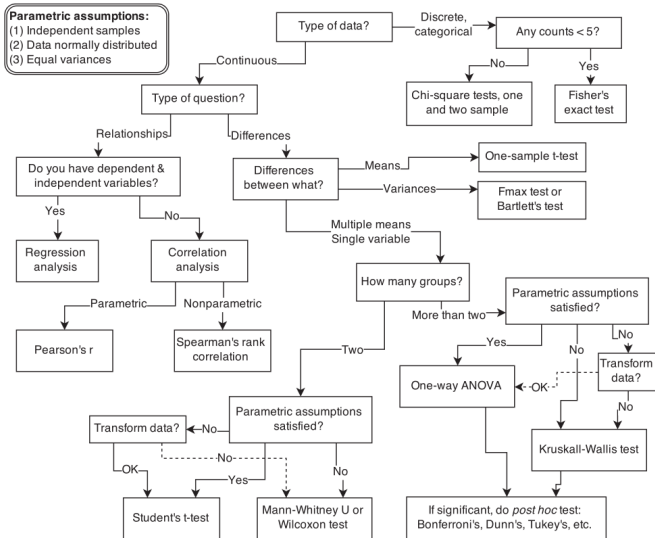


All is one. . .

...but `lm()` rules (IMHO)

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

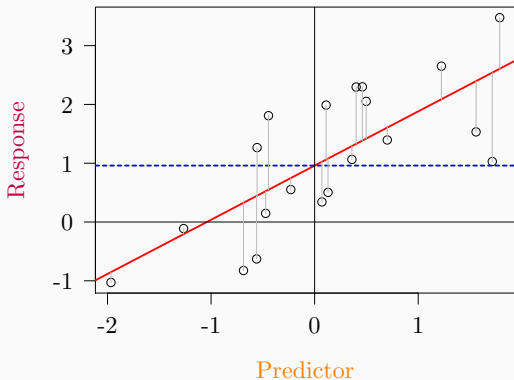
All is one...



ALL can be done as linear models

Focus on linear models

$$\text{Response} = \text{Intercept} + \text{Slope} \times \text{Predictor} + \text{Error}$$



A simple linear model

$$\text{Response} = \text{Intercept} + \text{Slope} \times \text{Predictor} + \text{Error}$$

```
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

Cautionary tales from the front

Introduction to Statistical Modelling

Another look at essential steps

General approach

1. Scientific question

General approach

1. Scientific question

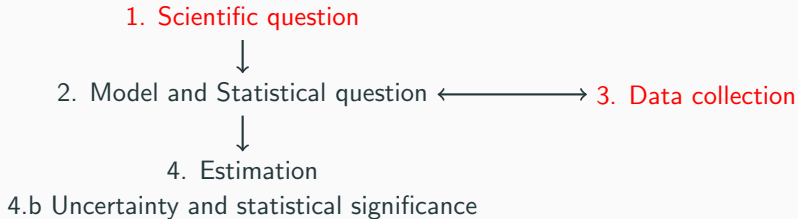


2. Model and Statistical question

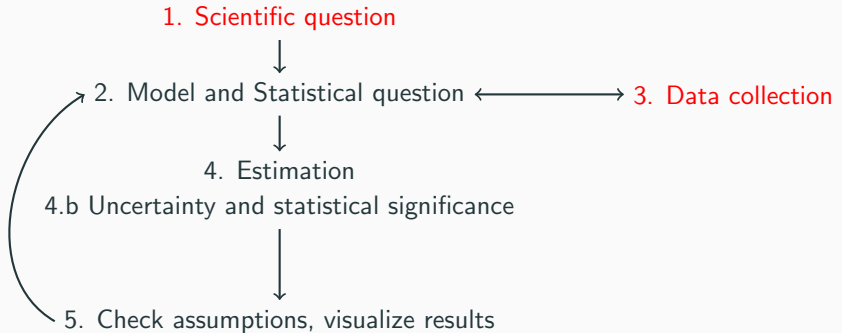
General approach



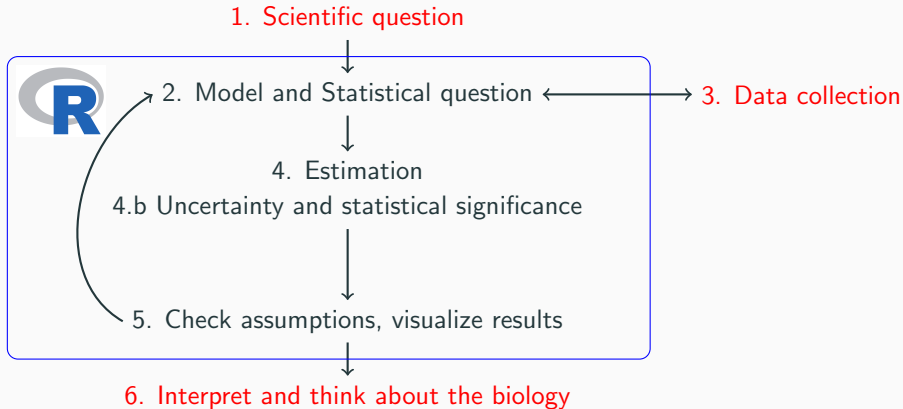
General approach



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:

$$\text{response} = A + D \times \text{predictor} + \epsilon$$

$A = ?$, $D = ?$

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:

$$\text{response} = A + D \times \text{predictor} + \epsilon$$

$$A = 2.81857, D = 0.44235$$

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:

$$\text{response} = A + D \times \text{predictor} + \epsilon$$

$$A = 2.81857, D = 0.44235$$

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:

$$\text{response} = A + D \times \text{predictor} + \epsilon$$

$$A = 2.81857, D = 0.44235$$

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)
```

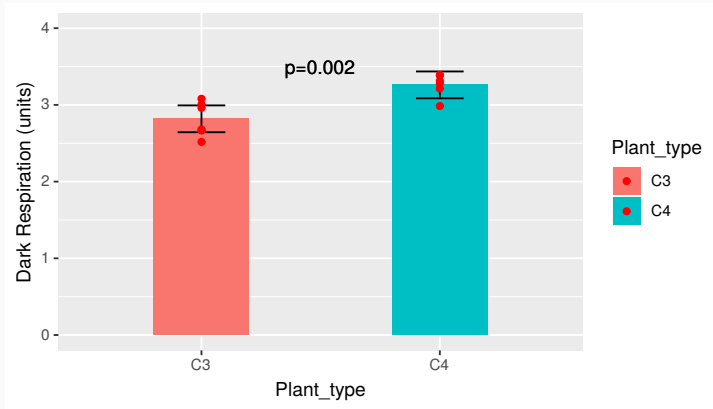
Check assumptions, visualize results

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



Another example

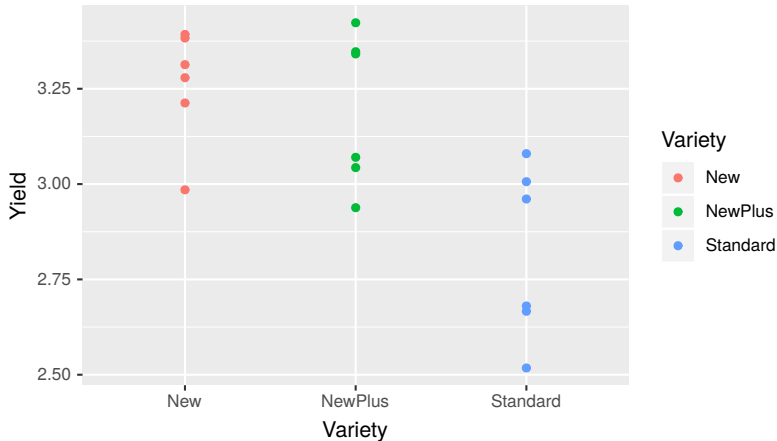
Compare wheat yields between 3 varieties

- Outcome measure: Tonnes/hectare
- Experimental factor: Variety (new/newPLUS/standard)
- Data: 6 plots/ variety



How many parameters in this model?

Results from Wheat Yield Experiment with 3 Varieties



Analyse these data in R

```
wheat2<-read.csv("wheat yield PLUS.csv")  
str(wheat2) #check data types for each variable  
View(wheat2) #View data  
ggplot(wheat2, aes(Variety, Yield, colour=Variety)) +  
  geom_point()
```


Sample analysis in R: 1-way ANOVA

```
aov1=aov(Yield~Variety, data = wheat2)
summary(aov1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Variety	2	0.6820	0.3410	8.951	0.00276 **
Residuals	15	0.5714	0.0381		

ANOVA

- Compares means between TWO or MORE GROUPS
- Relies on F-statistic = $\frac{\text{Between-groups variance}}{\text{Within groups variance}} = \frac{\text{Explained}}{\text{Unexplained}}$
- One test for significance of all groups

Sample analysis in R: 1-way ANOVA

```
aov1=aov(Yield~Variety, data = wheat2)
summary(aov1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Variety	2	0.6820	0.3410	8.951	0.00276 **
Residuals	15	0.5714	0.0381		

ANOVA

- Compares means between TWO or MORE GROUPS
- Relies on F-statistic = $\frac{\text{Between-groups variance}}{\text{Within groups variance}} = \frac{\text{Explained}}{\text{Unexplained}}$
- One test for significance of all groups

But which groups are different???

Using emmeans to extract group estimates

```
emmeans(aov1, ~Variety)
```

Variety	emmean	SE	df	lower.CL	upper.CL
Standard	2.81	0.079	15	2.64	2.98
New	3.26	0.079	15	3.09	3.43
NewPlus	3.19	0.079	15	3.02	3.36

Confidence level used: 0.95

Using emmeans to compare groups

```
emmeans(aov1, pairwise~Variety)
```

```
$contrasts
```

contrast	estimate	SE	df	t.ratio	p.value
Standard - New	-0.44	0.112	15	-3.925	0.0036
Standard { NewPlus	-0.37	0.112	15	-3.330	0.0120
New - NewPlus	0.06	0.112	15	0.595	0.8248

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

Equivalent with `lm()`

```
lm2<-lm(Yield ~ Variety, data = wheat2)
anova(lm2)
summary(lm2)
```

Analysis of Variance Table

Response: Yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Variety	2	0.68203	0.34101	8.9513	0.002764 **
Residuals	15	0.57145	0.03810		

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.81857	0.07968	35.372	7.26e-16 ***
VarietyNew	0.44235	0.11269	3.925	0.00135 **
VarietyNewPlus	0.37529	0.11269	3.330	0.00457 **

Equivalent with `lm()`

```
emmeans(lm2, pairwise~Variety)
```

\$emmeans

Variety	emmean	SE	df	lower.CL	upper.CL
Standard	2.81	0.0796	15	2.64	2.98
New	3.26	0.0796	15	3.09	3.43
NewPlus	3.19	0.0796	15	3.02	3.36

Confidence level used: 0.95

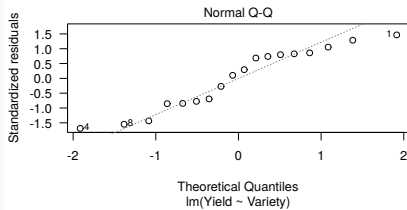
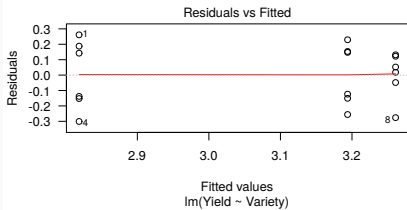
\$contrasts

contrast	estimate	SE	df	t.ratio	p.value
Standard - New	-0.44	0.112	15	-3.925	0.0036
Standard - NewPlus	-0.37	0.112	15	-3.330	0.0120
New - NewPlus	0.067	0.112	15	0.595	0.8248

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

Assessing model assumptions for `lm()`

```
plot(lm2)
```



Summary of results with post hoc comparisons:

