Laboratory & Professional Skills:
Data Analysis

Emma Rand
Data Analysis in R

More than one explanatory variable: Two-way ANOVA

Last week

- Extended our ability to test for differences between two or more groups with the one-way ANOVA and its non-parametric equivalent Kruskal-Wallis
  - Rationale for ANOVA rather than multiple *t*-tests
  - ANOVA terminology and concepts: SS, MS, *F*
  - Post-hoc tests
- In RStudio
  - Importing data from csv files
  - One way ANOVA and Kruskal-Wallis and their post-hoc tests
  - Summarising and reporting
  - Figure annotations

Summary of this week

Extend of our understanding to testing with two explanatory variables using the two-way ANOVA

- Comparison with one-way ANOVA
- The three null hypotheses
- Running, interpreting and reporting a two-way ANOVA
- Investigating the assumptions
- Understanding the interaction
- Post-hoc analysis (after a significant two-way ANOVA)

Learning objectives for the week

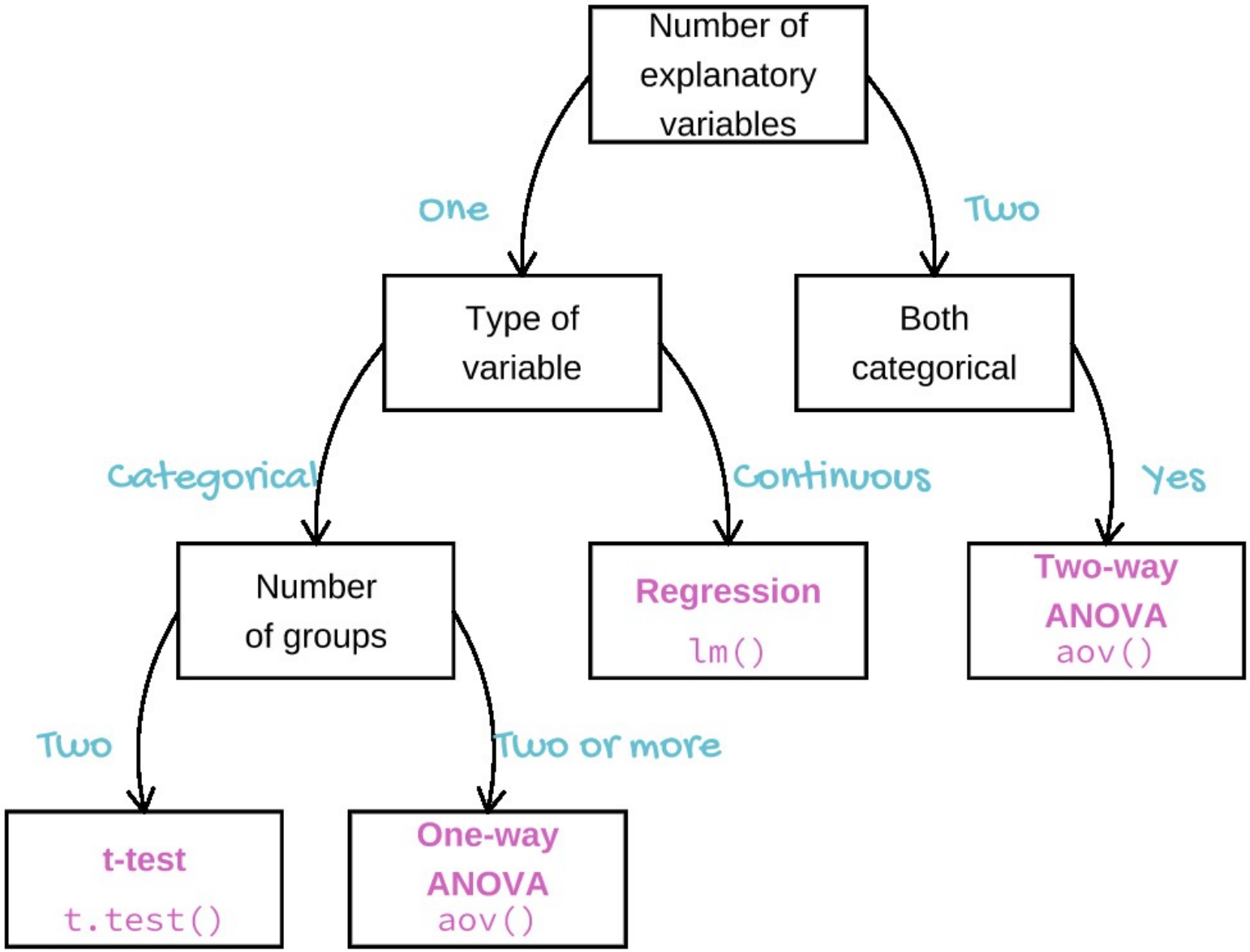
By the end of this week the successful student should be able to:

- List the sheets in, and read data from, Excel files (MLO 3)
- Combine dataframes of the same structure (MLO 3)
- Select, appropriately two-way ANOVA (MLO 2)
- Understand the meaning of the interaction term (MLO 2)
- Apply and interpret the two-way ANOVA (including the interaction), and post-hoc tests in R (MLO 3 and 4)
- Evaluate whether the assumptions of the test are met (MLO 2)
- Summarise and illustrate with appropriate R figures test results scientifically (MLO 3 and 4)

Review and rationale

Choosing tests: 3 steps

1. What is a one sentence description of what you want to know?
2. What are your explanatory variables?
   - Categories: *t*-tests, ANOVA, Wilcoxon, Mann-Whitney
   - Continuous: Regression, correlation
3. What is your response variable?
   - Normally distributed: *t*-tests, ANOVA, regression
   - Counts: Chi-squared or stage 2 ☺


Number of explanatory variables
One
Two
Type of variable
Both categorical
Categorical
Continuous
Yes
Number of groups
Regression
lm()
Two-way ANOVA
aov()
Two
Two or more
t-test
t.test()
One-way ANOVA
aov()

Choosing between one-way and two-way ANOVA?

Response:
wing lengths

Explanatory:
species

Untitled1* × butter ×

Filter

	winglen	spp
1	23.6	F.flappa
2	23.3	F.flappa
3	18.2	F.flappa
4	22.6	F.flappa
5	29.3	F.flappa
6	22.2	F.flappa
7	24.5	F.flappa
8	26.3	F.flappa
9	20.6	F.flappa
10	23.9	F.flappa
11	26.5	F.flappa
12	24.7	F.flappa
13	28.3	F.flappa
14	22.3	F.flappa
15	21.8	F.flappa
16	30.0	F.flappa
17	21.5	F.flappa
18	20.1	F.flappa
19	24.3	F.flappa
20	27.2	F.flappa
21	28.6	F.concocti
22	17.2	F.concocti
23	20.4	F.concocti
24	21.9	F.concocti
25	26.3	F.concocti
26	27.8	F.concocti

Showing 1 to 26 of 40 entries

Choosing between one-way and two-way ANOVA?

What if we have two explanatory variables?

- Two one-way ANOVAs?? NO
- A Two-way ANOVA YES

- Note: tidy data format


lect 080 two-way ANOVA.R × butter ×

Filter

	winglen	spp	region
1	23.6	F.flappa	south
2	23.3	F.flappa	south
3	18.2	F.flappa	south
4	22.6	F.flappa	south
5	29.3	F.flappa	south
6	22.2		
7	24.5		
8	26.3		
9	20.6		
10	23.9		
11	26.5		
12	24.7	F.flappa	north
13	28.3	F.flappa	north
14	22.3	F.flappa	north
15	21.8	F.flappa	north
16	30.0	F.flappa	north
17	21.5	F.flappa	north
18	20.1	F.flappa	north
19	24.3	F.flappa	north
20	27.2	F.flappa	north
21	28.6	F.concocti	south
22	17.2	F.concocti	south
23	20.4	F.concocti	south
24	21.9	F.concocti	south

Showing 1 to 24 of 40 entries, 3 total columns

Explanatory:
species
region

Two-way ANOVA

Assumptions

- Same as for one-way ANOVA
- Normality and homogeneity of variance in residuals
  - Common sense
  - **Check after ANOVA** using the $residuals variable and diagnostic plots (as we did after one-way ANOVA)

The two-way ANOVA

A parametric test

Two-way ANOVA

Example

Response: wing lengths

Explanatory variables:

region: two levels

spp: two levels

lect 080 two-way ANOVA.R × | butter ×

Filter

	winglen	spp	region
1	23.6	F.flappa	south
2	23.3	F.flappa	south
3	18.2	F.flappa	south
4	22.6	F.flappa	south
5	29.3	F.flappa	south
6	22.2	F.flappa	south
7	24.5	F.flappa	south
8	26.3	F.flappa	south
9	20.6	F.flappa	south
10	23.9	F.flappa	south
11	26.5	F.flappa	north
12	24.7	F.flappa	north
13	28.3	F.flappa	north
14	22.3	F.flappa	north
15	21.8	F.flappa	north
16	30.0	F.flappa	north
17	21.5	F.flappa	north
18	20.1	F.flappa	north
19	24.3	F.flappa	north
20	27.2	F.flappa	north
21	28.6	F.concocti	south
22	17.2	F.concocti	south
23	20.4	F.concocti	south
24	21.9	F.concocti	south

Showing 1 to 24 of 40 entries, 3 total columns

Two-way ANOVA example

What does it test?

The null hypotheses here are:

1. mean of *F.flappa* (averaged over the regions) = mean of *F.concocti* (averaged over the regions),
2. mean of north (averaged over the spp) = mean of south (averaged over the spp) and
3. the effects of the two factors are independent.

Two-way ANOVA example

What does it test?

The null hypotheses here are:

1. mean of *F.flappa* (averaged over the regions) = mean of *F.concocti* (averaged over the regions),
2. mean of north (averaged over the spp) = mean of south (averaged over the spp) and
3. the effects of the two factors are independent.

Two-way ANOVA example

Reading in and examining the structure of the data

```
butter <- read_table2("../data/butterf.txt")
glimpse(butter)
Observations: 40
Variables: 3
$ winglen <dbl> 23.6, 23.3, 18.2, 22.6, 29.3, 22.2, 24.5, 26.3, 20.6, 23.9...
$ spp     <fct> F.flappa, F.flappa, F.flappa, F.flappa, F.flappa, F.flappa...
$ region  <fct> south, south, south, south, south, south, south, south, so...
```

Assumptions

Common sense

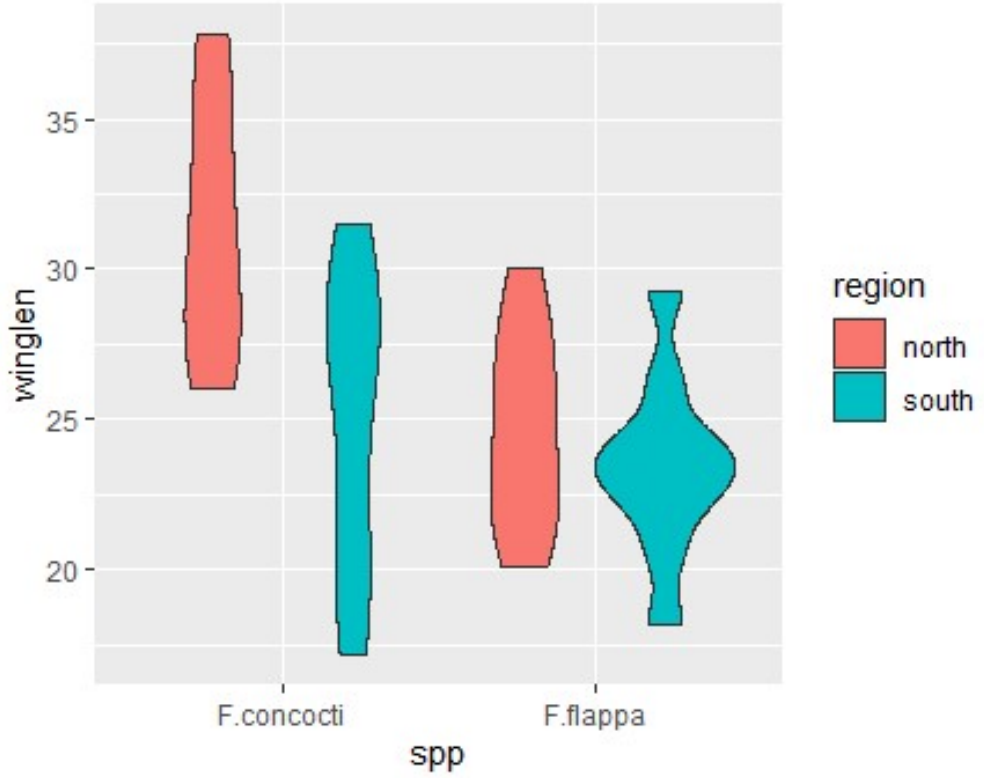
Can be checked after analysis

Two-way ANOVA example

Plot your data

Plot your data: roughly – perhaps..

```
ggplot(data = butter,
       aes(x = spp, y = winglen, fill = region)) +
  geom_violin()
```



Two-way ANOVA example

Example

Sumarise

```
buttersum <- butter %>%
  group_by(region, spp) %>%
  summarise(mean = mean(winglen),
            median = median(winglen),
            sd = sd(winglen),
            n = length(winglen),
            se = sd/sqrt(n))
```

```
buttersum
# A tibble: 4 x 7
# Groups:   region [2]
  region spp          mean median    sd     n    se
  <fct>  <fct>       <dbl>  <dbl> <dbl> <int> <dbl>
1 north  F.concocti  31.4   31.0  4.28    10 1.35
2 north  F.flappa    24.7   24.5  3.27    10 1.03
3 south  F.concocti  25.0   27.0  4.96    10 1.57
4 south  F.flappa    23.4   23.5  3.01    10 0.953
```

Example

Run the ANOVA

```
mod <- aov(data = butter,
           winglen ~ region * spp)
```

Assign result because we will be able to access residuals from this object later

Two-way ANOVA example

Example

Run the anova

```
mod <- aov(data = butter,
           winglen ~ region * spp)
```

The model: explain winglen by region, spp and the interaction between them

Two-way ANOVA example

Understanding the test output

```
mod <- aov(data = butter, winglen ~ region * spp)
summary(mod)
             Df Sum Sq Mean Sq F value   Pr(>F)
region        1 145.16 145.161  9.2717 0.004334 **
spp           1 168.92 168.921 10.7893 0.002280 **
region:spp    1  67.08  67.081  4.2846 0.045692 *
Residuals    36 563.63  15.656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Three tests
Three p values

Two-way ANOVA example

Understanding the test output

```
mod <- aov(data = butter, winglen ~ region * spp)
summary(mod)
             Df Sum Sq Mean Sq F value   Pr(>F)
region        1 145.16 145.161  9.2717 0.004334 **
spp           1 168.92 168.921 10.7893 0.002280 **
region:spp    1  67.08  67.081  4.2846 0.045692 *
Residuals    36 563.63  15.656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1. There is an effect of region (difference between regions)
2. There is an effect of species (difference between species)
3. There is an interaction between region and species…..

Two-way ANOVA example

Understanding the test output

```
mod <- aov(data = butter, winglen ~ region * spp)
summary(mod)
             Df Sum Sq Mean Sq F value   Pr(>F)
region        1 145.16 145.161  9.2717 0.004334 **
spp           1 168.92 168.921 10.7893 0.002280 **
region:spp    1  67.08  67.081  4.2846 0.045692 *
Residuals    36 563.63  15.656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Total d.f. is no. of values – 1: 40 – 1 = 39
region d.f. is no. regions – 1: 2 – 1 = 1
spp d.f. is no. spp – 1: 2 – 1 = 1
Interaction d.f. is region d.f. * spp d.f. : 1* 1 = 1
Residual d.f. is total d.f. – all other d.f.: 39 – 1 – 1 – 1 = 36

Two-way ANOVA example

Understanding the test output

```
mod <- aov(data = butter, winglen ~ region * spp)
summary(mod)
             Df Sum Sq Mean Sq F value   Pr(>F)
region        1 145.16 145.161  9.2717 0.004334 **
spp           1 168.92 168.921 10.7893 0.002280 **
region:spp    1  67.08  67.081  4.2846 0.045692 *
Residuals    36 563.63  15.656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

‘Error term’ for all 3 tests

Two-way ANOVA example

Checking Assumptions

- Common sense
  - response should be continuous
  - No/few repeats
- Plot the residuals
- Using a test in R

Two-way ANOVA

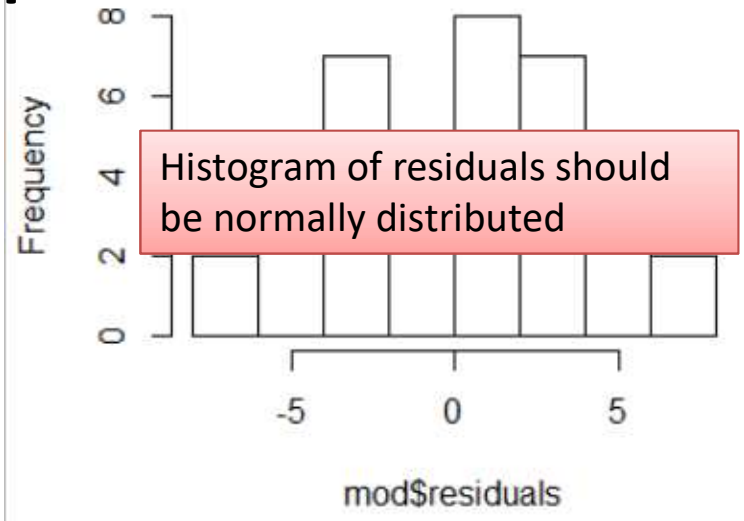
Checking Assumptions

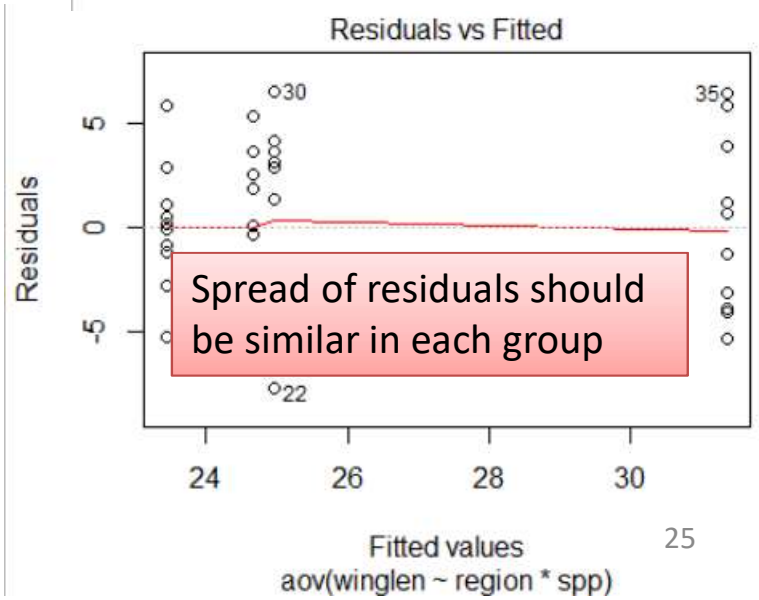
Residuals are calculated for you already!

```
hist(mod$residuals)
shapiro.test(mod$residuals)

   Shapiro-Wilk normality test

data:  mod$residuals
W = 0.97306, p-value = 0.4474
plot(mod, which=1)
```





Two-way ANOVA example

Reporting the result

Reporting the result: “significance, direction, magnitude”

There was a significant difference between the species (ANOVA: $F = 10.79$; $d.f. = 1,36$; $p = 0.002$) and between the regions ($F = 9.27$; $d.f. = 1,36$; $p = 0.004$). However, there was also a significant interaction between region and species ($F = 4.28$; $d.f. = 1,36$; $p = 0.046$)

What about direction and magnitude??

Two-way ANOVA example

Reporting the result: Post-hoc?

Post-hoc test e.g., Tukey

John Wilder Tukey

Wild Turkey

Wild Turkey

Two-way ANOVA example

Reporting the result

Which means differ? Post-hoc test needed e.g., Tukey

3 parts to the output. First two parts for region and spp

```
TukeyHSD(mod)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = winglen ~ region * spp, data = butter)

$region
              diff       lwr       upr    p adj
south-north -3.81 -6.347658 -1.272342 0.004334

$spp
                      diff       lwr       upr     p adj
F.flappa-F.concocti -4.11 -6.647658 -1.572342 0.0022796
```

Two-way ANOVA example

Reporting the result

Which means differ? Post-hoc test needed e.g., Tukey

3 parts to the output. Third part for the interaction

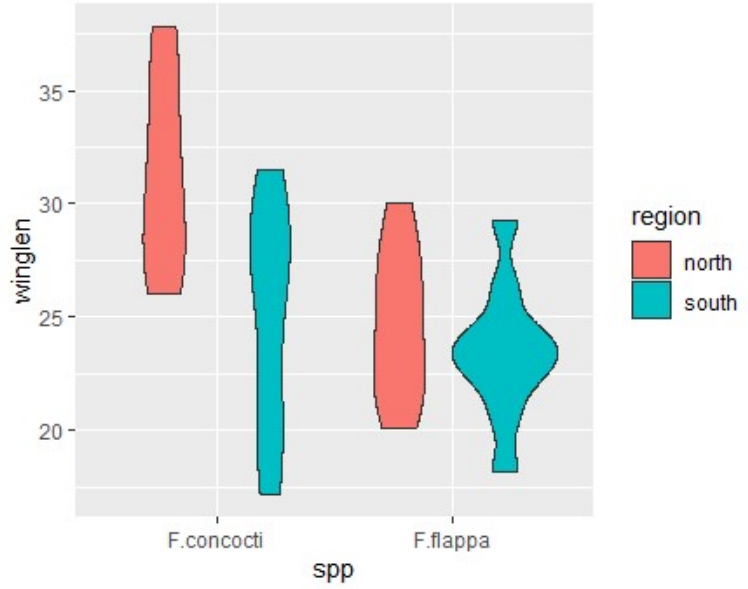
```
$`region:spp`
                                      diff        lwr       upr     p adj
south:F.concocti-north:F.concocti -6.40 -11.165769 -1.634231 0.0048102
north:F.flappa-north:F.concocti   -6.70 -11.465769 -1.934231 0.0030099
south:F.flappa-north:F.concocti   -7.92 -12.685769 -3.154231 0.0004123
north:F.flappa-south:F.concocti   -0.30  -5.065769  4.465769 0.9982343
south:F.flappa-south:F.concocti   -1.52  -6.285769  3.245769 0.8257284
south:F.flappa-north:F.flappa     -1.22  -5.985769  3.545769 0.9004525
```

Two-way ANOVA example

Reporting the result: direction and magnitude

```
$`region:spp`
                                        diff        lwr       upr     p adj
south:F.concocti-north:F.concocti -6.40 -11.165769 -1.634231 0.0048102
north:F.flappa-north:F.concocti   -6.70 -11.465769 -1.934231 0.0030099
south:F.flappa-north:F.concocti   -7.92 -12.685769 -3.154231 0.0004123
north:F.flappa-south:F.concocti   -0.30  -5.065769  4.465769 0.9982343
south:F.flappa-south:F.concocti   -1.52  -6.285769  3.245769 0.8257284
south:F.flappa-north:F.flappa     -1.22  -5.985769  3.545769 0.9004525
```

```
buttersum
# A tibble: 4 x 7
# Groups:   region [2]
  region spp          mean
  <fct>  <fct>       <dbl>
1 north  F.concocti   31.4
2 north  F.flappa     24.7
3 south  F.concocti   25.0
4 south  F.flappa     23.4
```



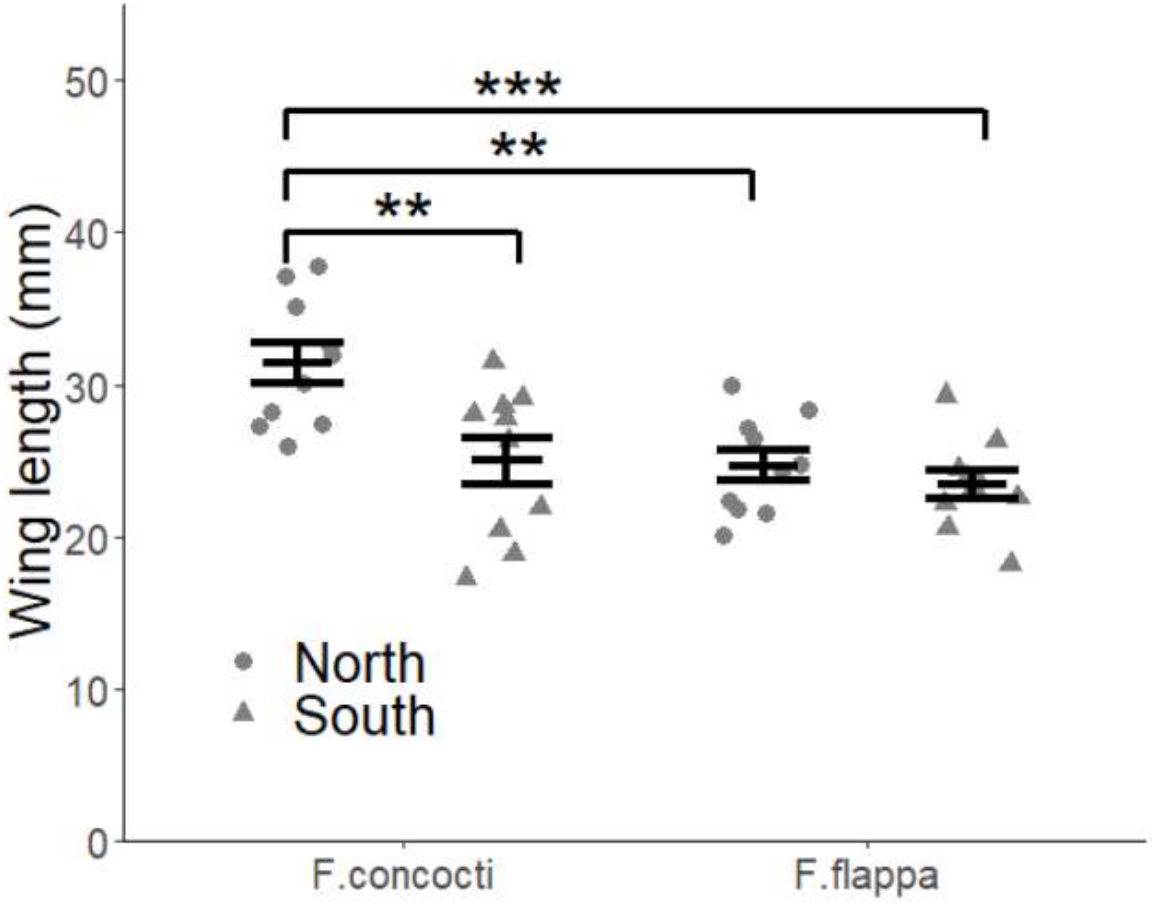
Two-way ANOVA example

Reporting the result: direction and magnitude

F.concocti had significantly longer wings on average than *F.flappa* (ANOVA: $F = 10.79$; *d.f*. = 1,36; $p = 0.002$) and individuals were significantly bigger in the North than the South ($F = 9.27$; *d.f.* = 1,36; $p = 0.004$). However, this was primarily because northern *F.concocti* were significantly larger than the other three groups (Tukey Honest Significant difference: southern *F.concocti* $p = 0.0048$, southern *F.flappa* $p = 0.0004$, northern *F.flappa* $p = 0.003$,). See Figure 1.

Two-way ANOVA example

Reporting the result: figure



Two-way ANOVA summary

- Parametric
- Two explanatory categorical variables
- Three null hypotheses: main effects and interaction
- Function in R:

```
mod <- aov(data = df, response ~ explanatory1 * explanatory2)
summary(mod)
```

- assumptions: normally and homogenously distributed residuals

continued

Two-way ANOVA summary

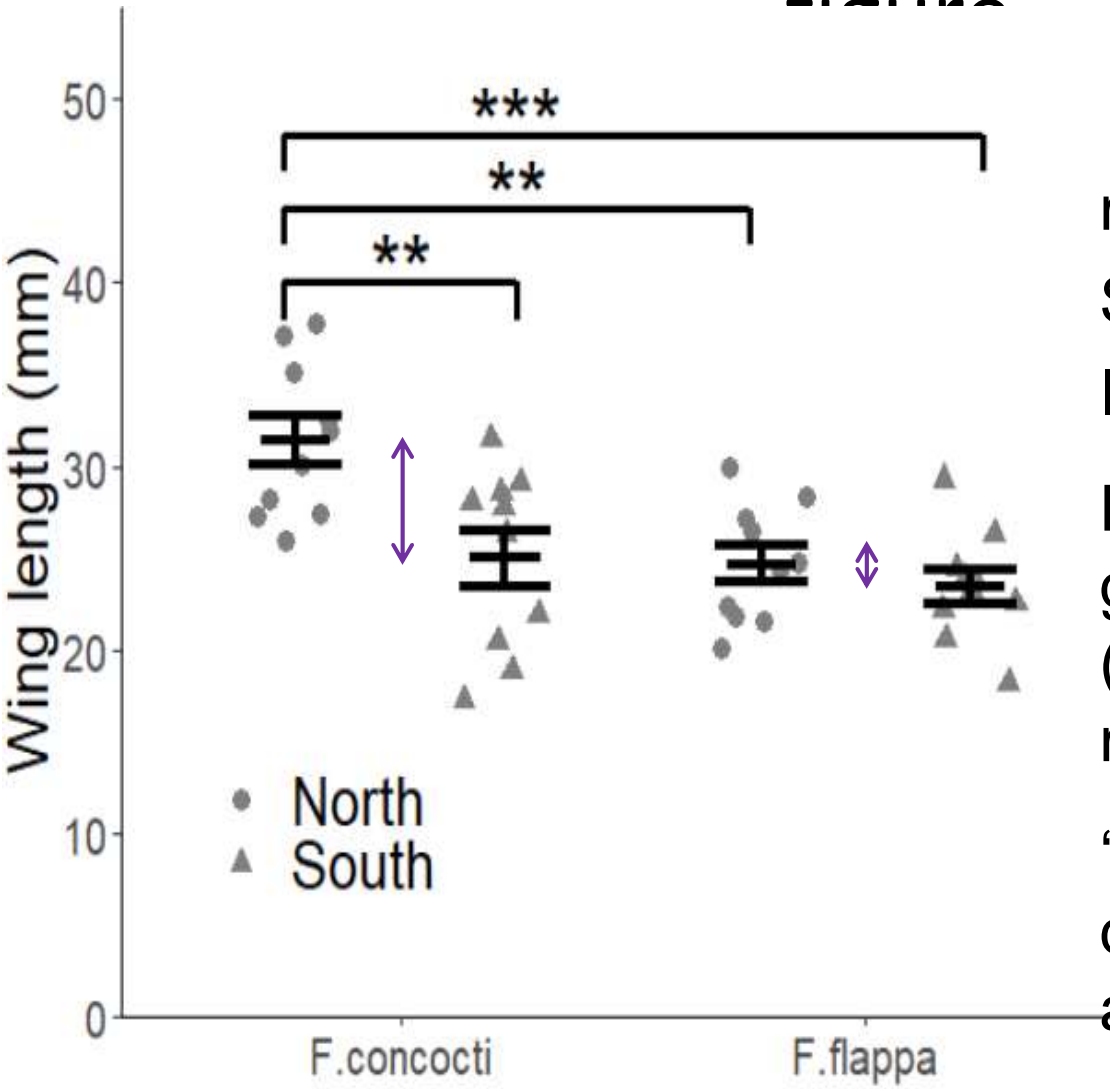
- ANOVA tells at lest two means differ and a post-hoc test is need to determine which means differ
- Tukey Honest Significant Difference is the post-hoc we used
- Function in R:
  `TukeyHSD(mod)`
- Interpret the interaction
- Significance, direction, magnitude
- Figure: data and ‘model’

Understanding the interaction

in two-way ANOVA

Two-way ANOVA example

Understanding the interaction from the figure



region – Sig

Spp – Sig

Int – Sig

Effect of region is greater in *F.concocti* (i.e., the gap between regions is bigger)

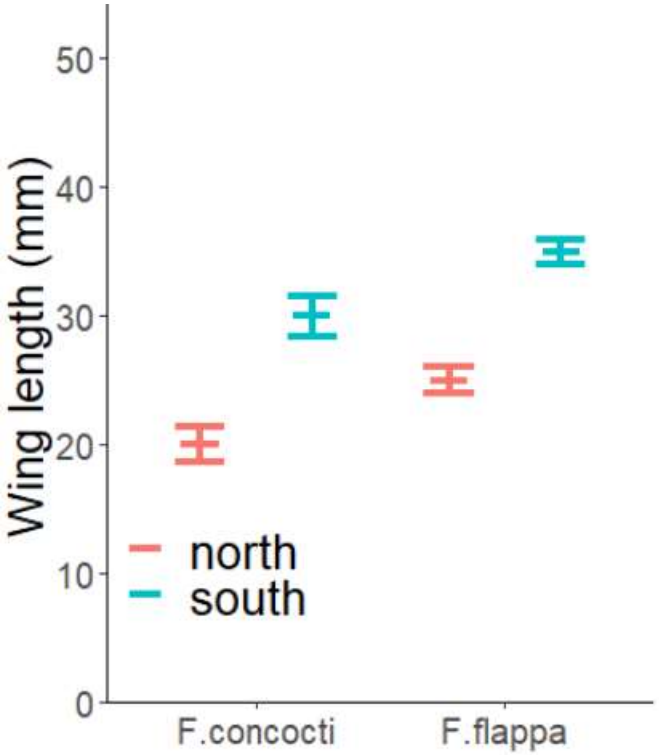
'effect of one factor depends on the level of another'

Two-way ANOVA example

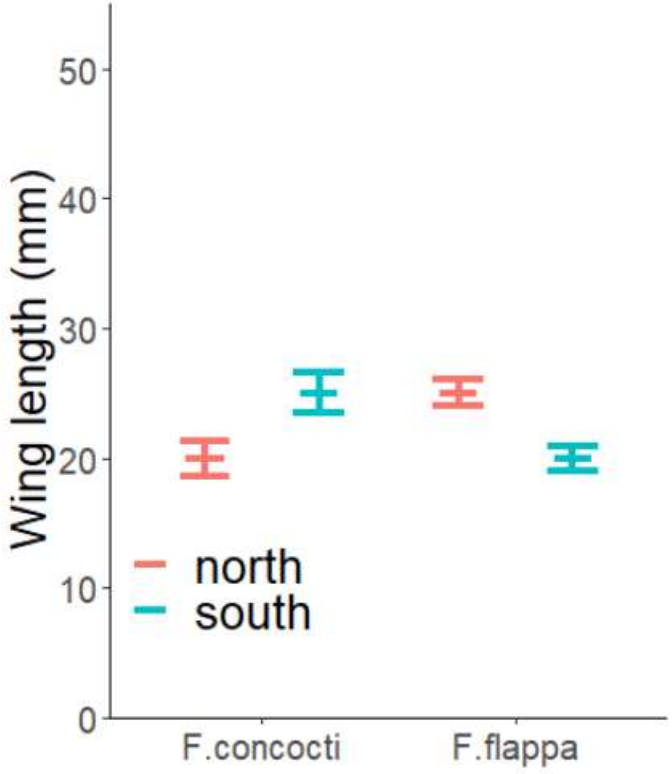
Understanding the interaction from the figure

Some other possible results

No interaction: Gap the same

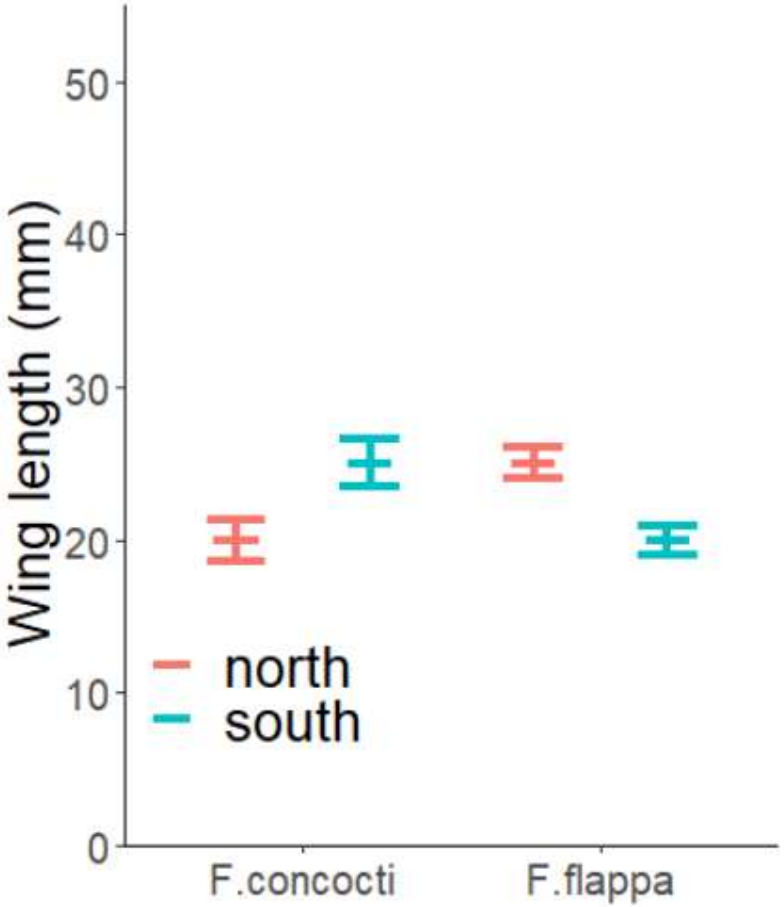


Interaction: Gap the reversed



Two-way ANOVA example

Understanding the interaction from the figure



Region – NS

Spp – NS

Int – Sig

But region does have an effect!

It is just reversed!

If you have a significant interaction, interpret main effects with care. Look at the Post-hoc test