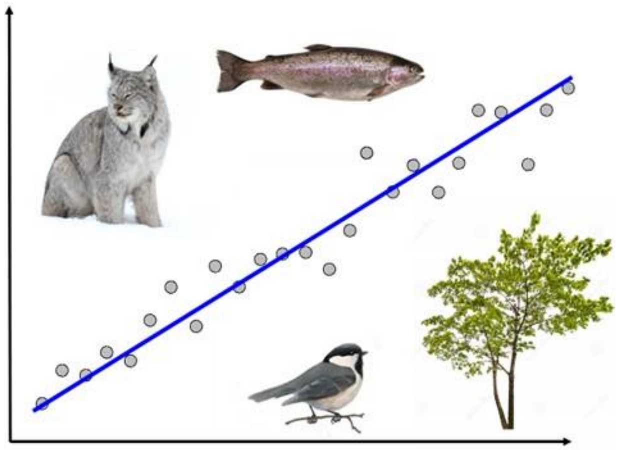
NRC 290b Introduction to Quantitative Ecology

Week 8 – Differences between more than two samples



Meg Graham MacLean, PhD

Department of Environmental Conservation

mgmaclean@umass.edu

2019 - Fall

This week

Monday

- Differences between many groups
  - One-way ANOVA
    - Sum of squares
  - Tukey HSD post-hoc test
  - Multi-way ANOVA

Wednesday

- Group exercise
  - Salamanders again! What about differences in lengths among 3 groups?

Why do we conduct difference tests?

Statistical testing

I am interested in testing whether there is a significant difference between the population sizes of fish in 30 low salinity lakes and 30 high salinity lakes.

Which statistical test should I use?

a) A t-test

b) A t-test for matched pairs

c) A one-way ANOVA

d) A two-way ANOVA

Statistical testing

I am interested in testing whether there is a significant difference between the population sizes of fish in 30 low salinity lakes and 30 high salinity lakes.

What is the test statistic for the test?

a) t

b) F

c) p

d) r

Statistical testing

I am interested in testing whether there is a significant difference between the population sizes of fish in 30 low salinity lakes and 30 high salinity lakes. In fact, I actually sampled 10 large, 10 medium, and 10 small lakes in each of the high and low salinity lakes. I want to explore whether there are differences in population size based on lake salinity and lake size.

Which statistical test should I use?

a) A t-test

b) A t-test for matched pairs

c) A one-way ANOVA

d) A two-way ANOVA

iclicker®

Statistical testing

I am interested in testing whether there is a significant difference between the population sizes of fish in 30 low salinity lakes and 30 high salinity lakes. In fact, I actually sampled 10 large, 10 medium, and 10 small lakes in each of the high and low salinity lakes. I want to explore whether there are differences in population size based on lake salinity and lake size.

What is the test statistic for the test?

a) t

b) F

c) p

d) r

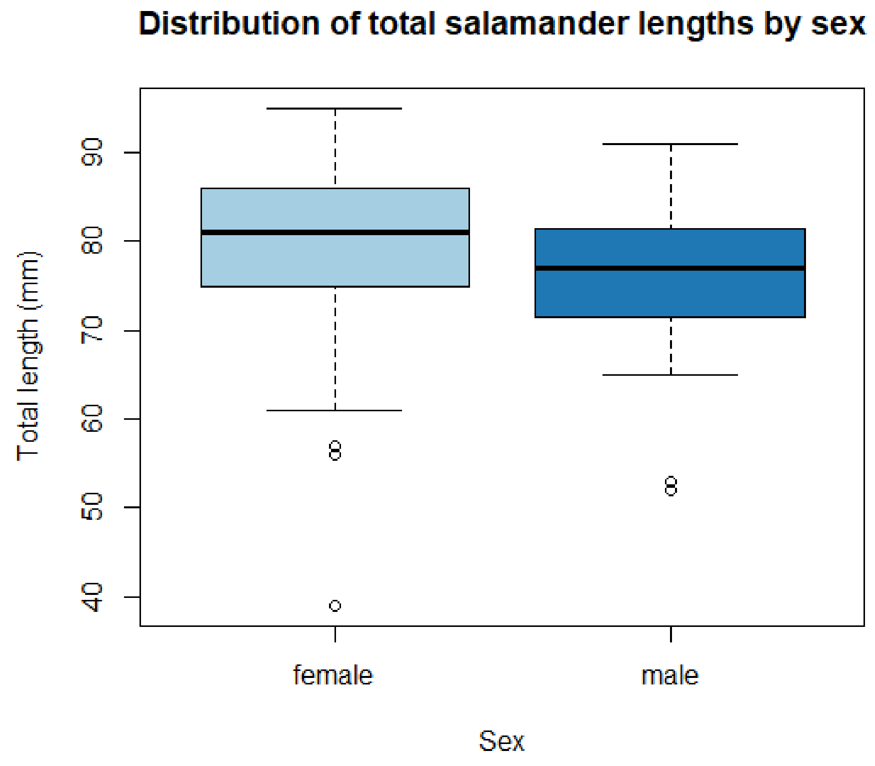
iclicker®

Differences

What do we do when we want to test the differences between two samples?

The t-test!

- H_0: there is *no difference* between means



Differences

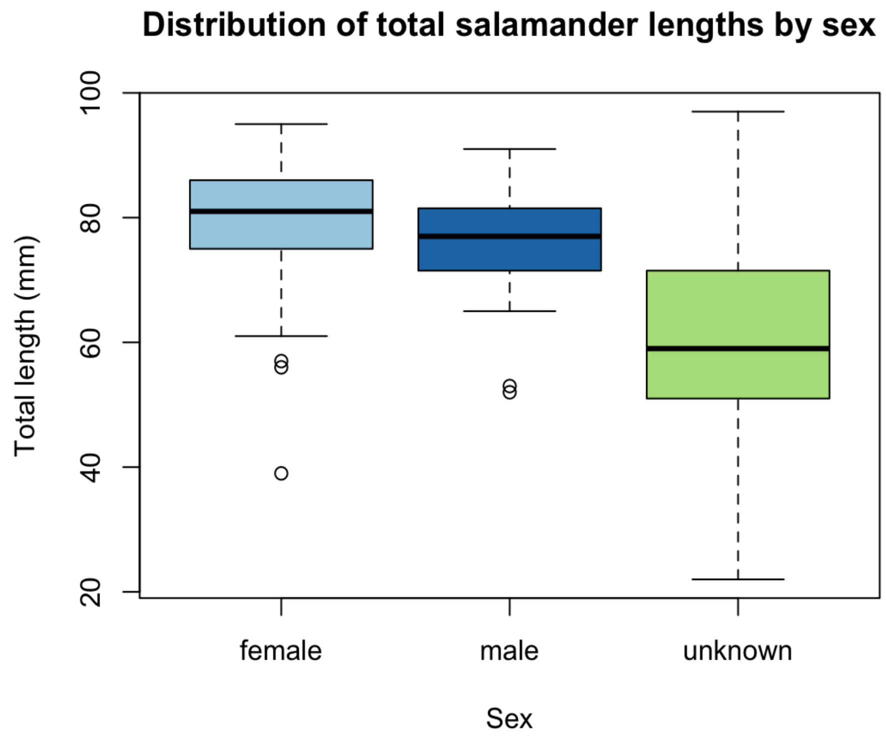
What do we do when we want to test the differences between two samples?

The t-test!

- H_0: there is *no difference* between means

But what if there are >2 samples/groups??

Analysis of Variance (ANOVA)



ANOVA example

Do plants' productivity (measured using biomass) differ with fertilizer treatment?

- Is there a difference between treatments?
- Is there a difference from the control (no fertilizer)?

ANOVA example

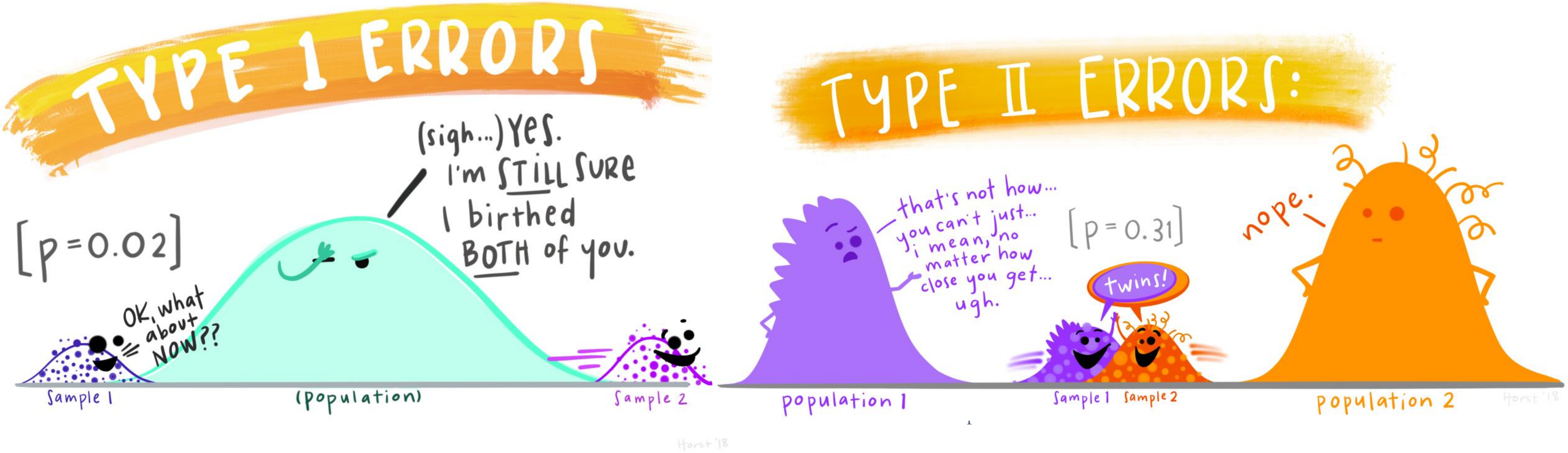
Why can't we use a t-test for checking differences?

- Need to do all possible pairs (very time consuming)
- Can get spurious differences by chance (type I error), since you are doing so many tests

ANOVA example

Why can't we use a t-test for checking differences?

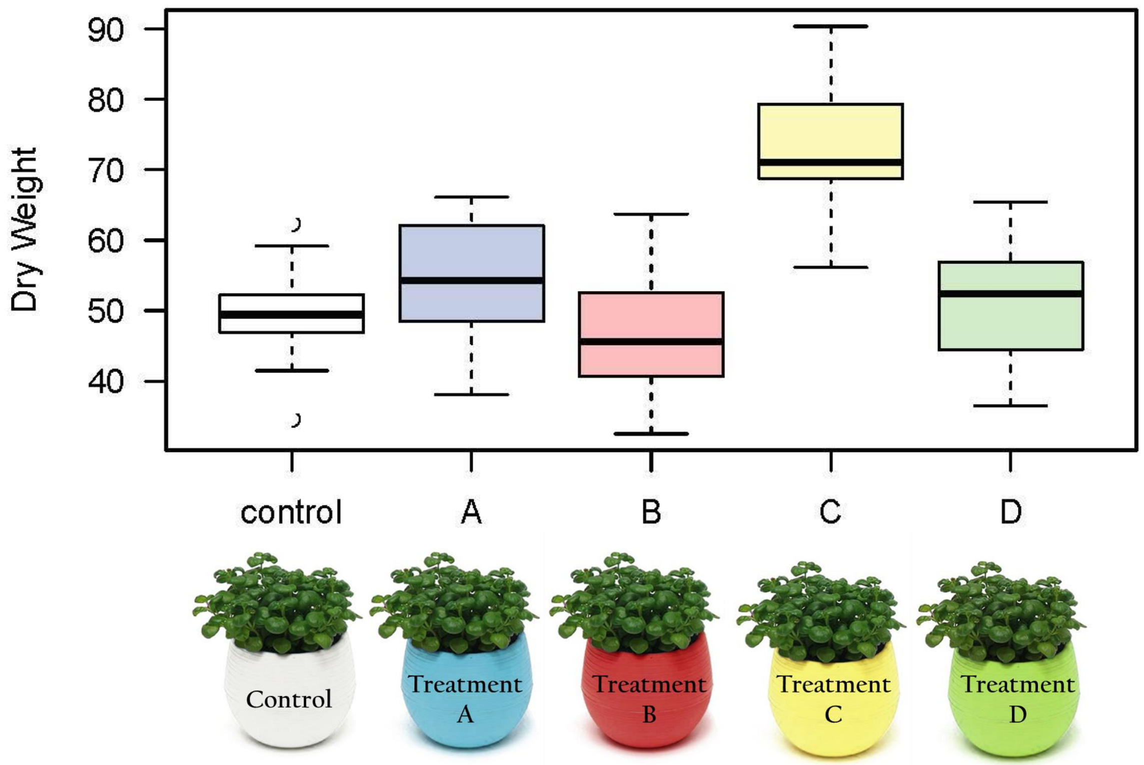
- Need to do all possible pairs (very time consuming)
- Can get spurious differences by chance (type I error), since you are doing so many tests



ANOVA example

ANOVA tests for significant difference among >2 groups

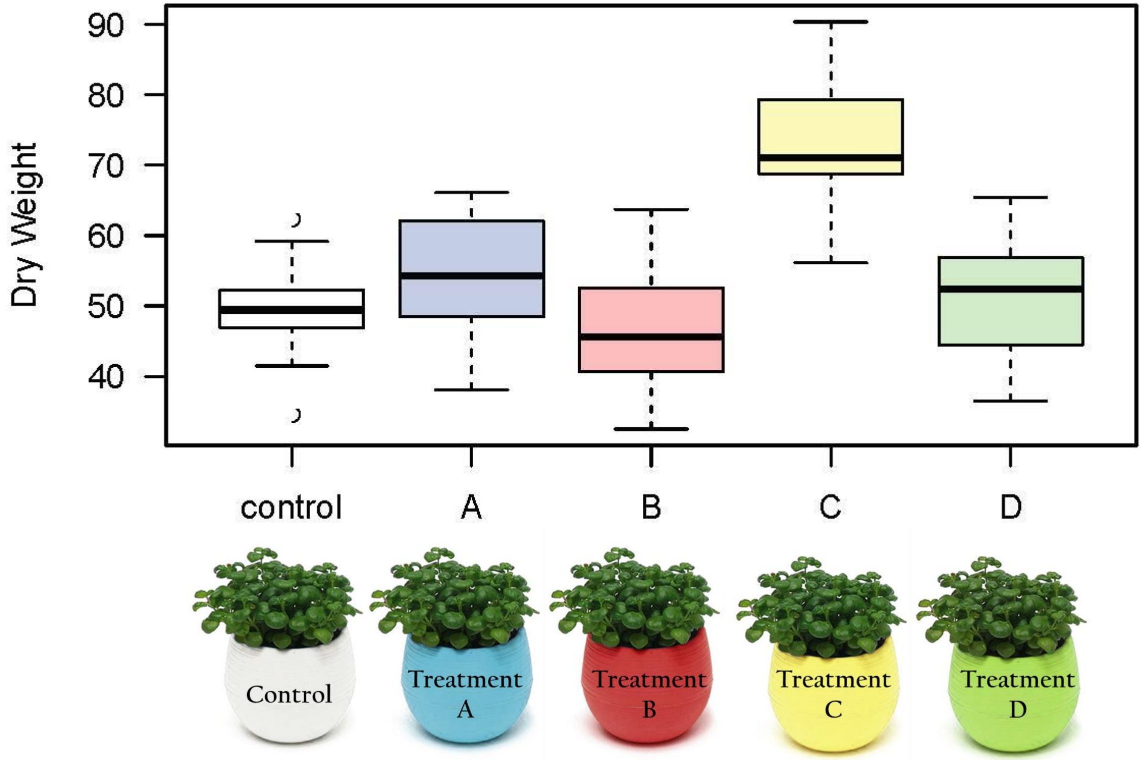
- Assumption: data are normally distributed
- Note: ANOVA and t-test are the same with 2 groups!



ANOVA example

ANOVA hypotheses

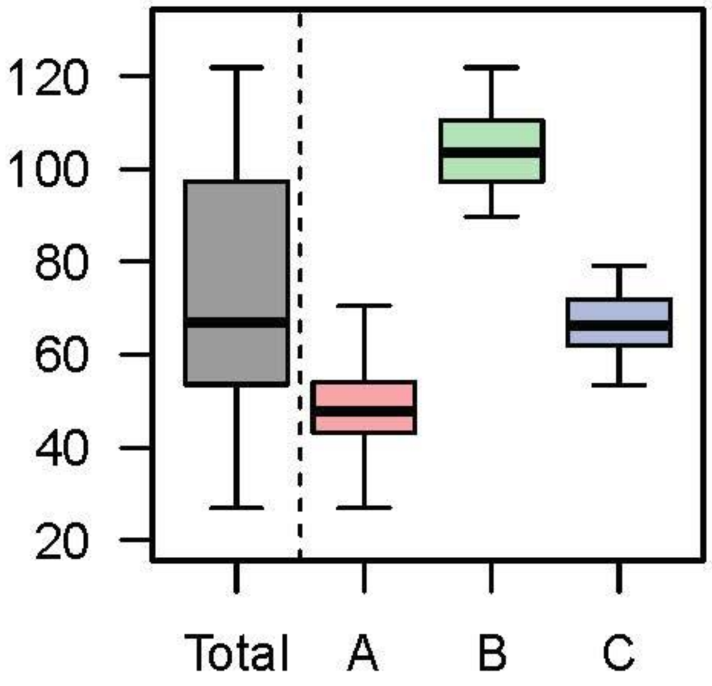
- H_0: there are no significant differences between means (i.e. all means equal)
- H_1: there are significant differences between means (i.e. *all* means not equal)



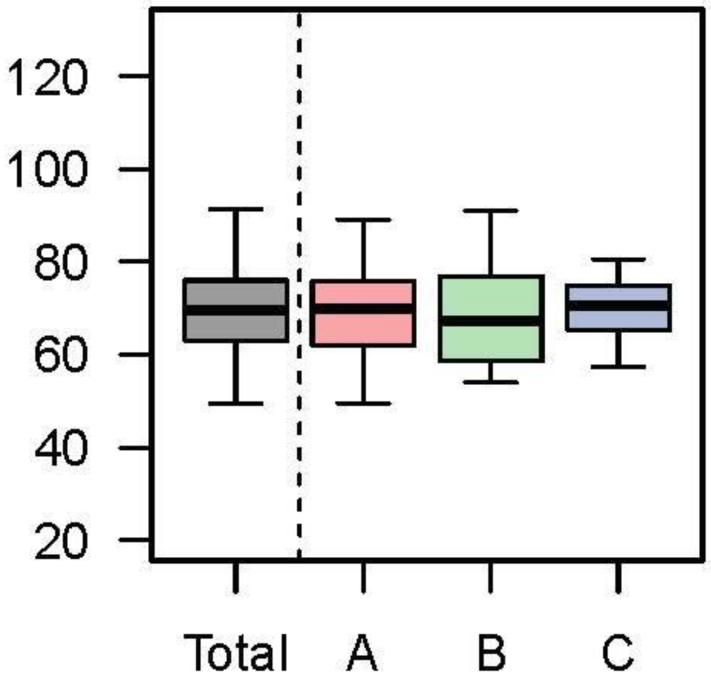
ANOVA

ANOVA test partitions the *total* variation into *within* sample variation and *between* sample variation to determine if samples come from a single distribution

High total variance



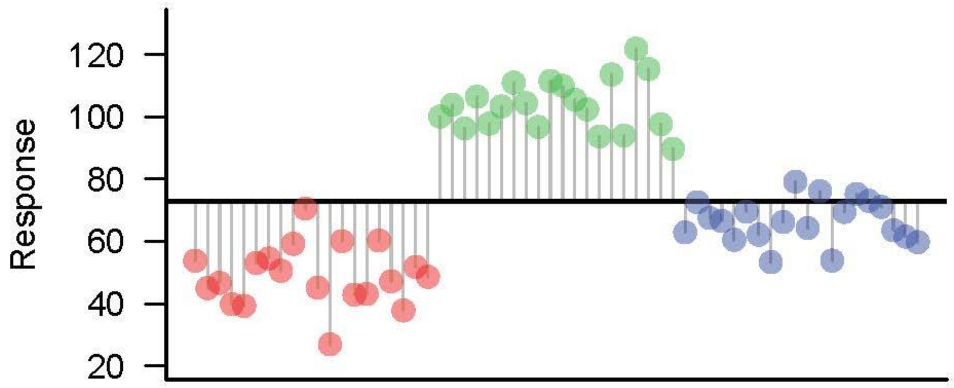
Low total variance

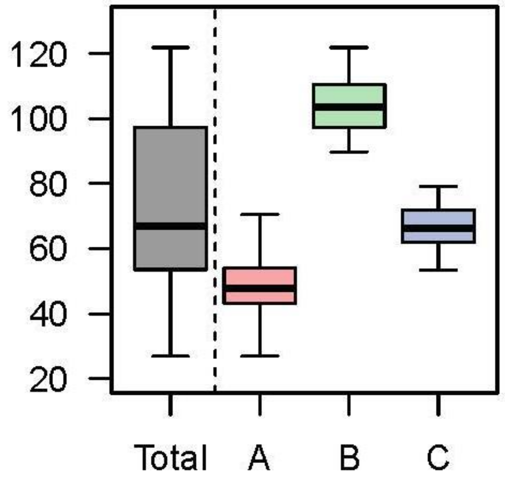


ANOVA – partitioning the variance

Sum of Squares (SS)!

- Total Sums of Squares (SS_T)
- Within-sample Sums of Squares (SS_W)
- Between-sample Sums of Squares (SS_B)



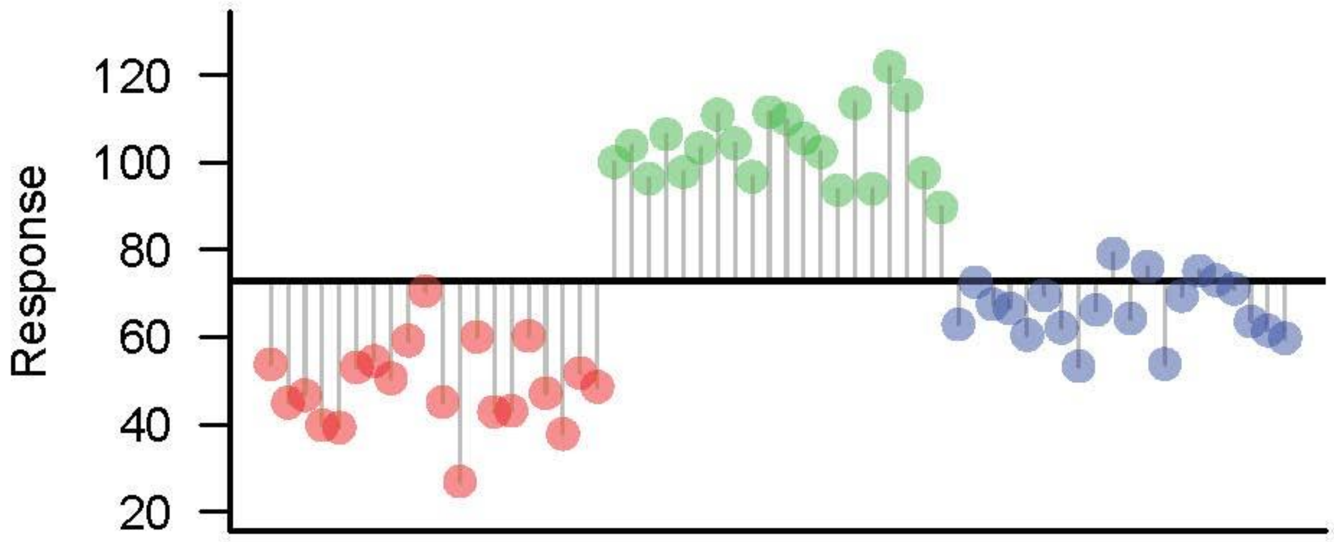


ANOVA – partitioning the variance

Total Sums of Squares (SS_T)

- The variability of the data from the *total* mean (not by group)
- Not (exactly) the variance! (not divided by the total number of samples)

$$SS_T = \sum (x - \bar{x})^2$$

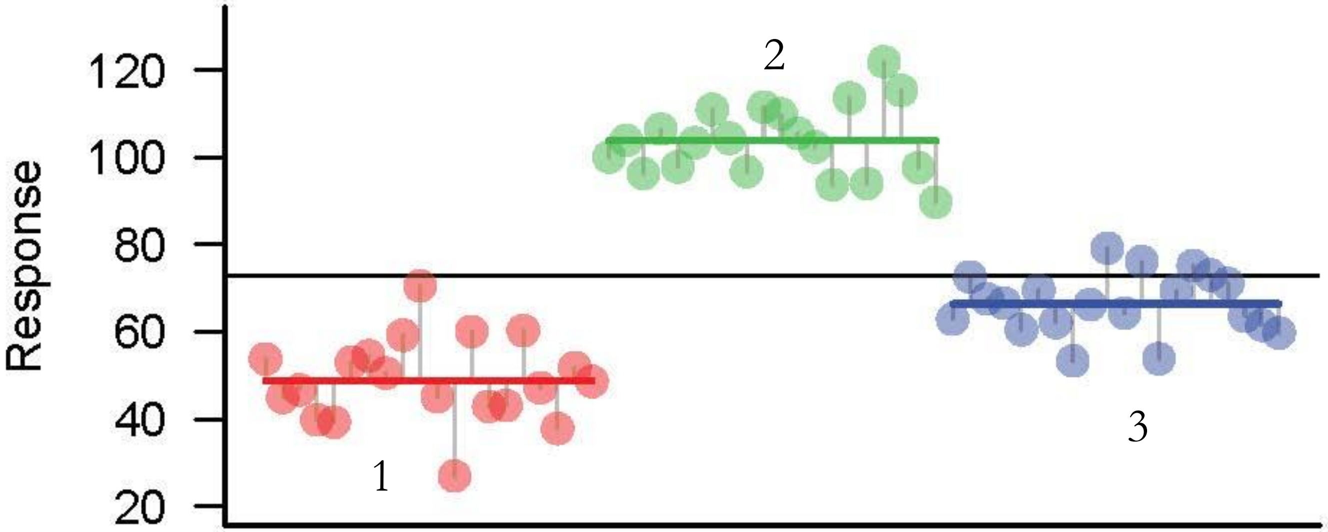


ANOVA – partitioning the variance

Within-sample Sums of Squares (SS_W)

- The variability of the data from the *group* mean

$$SS_W = \sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2 + \sum (x_3 - \bar{x}_3)^2$$

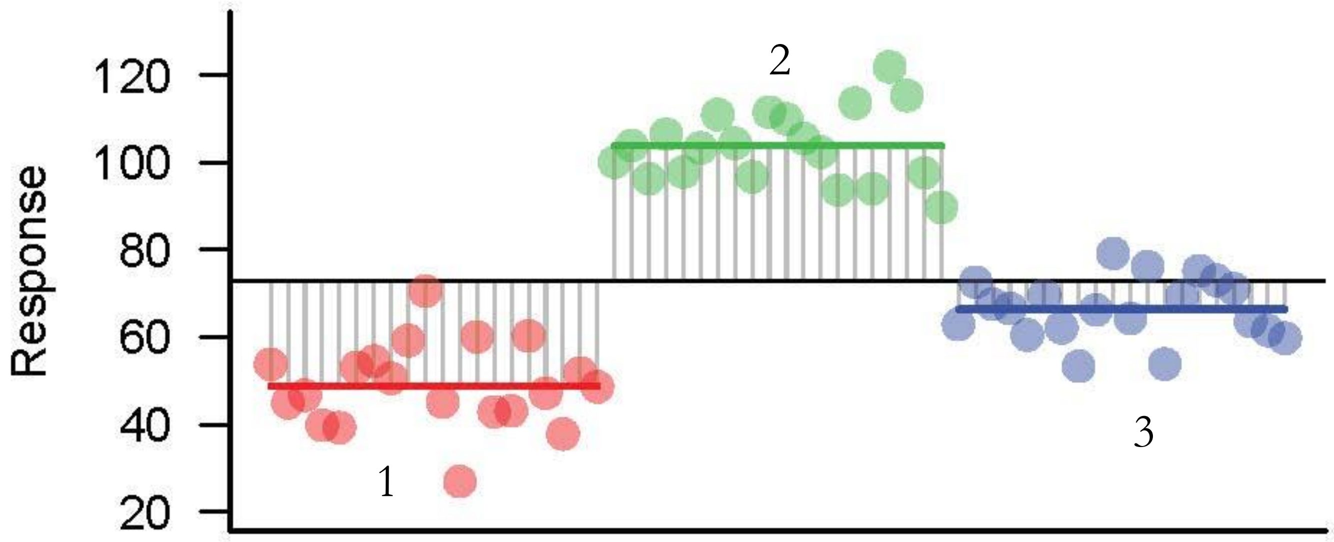


ANOVA – partitioning the variance

Between-sample Sums of Squares (SS_B)

- The variability of the *group* means from the *total* mean

$$SS_B = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2$$



Sum of Squares recap

$$SS_T = \sum (x - \bar{x})^2$$

- $df_T = n - 1$

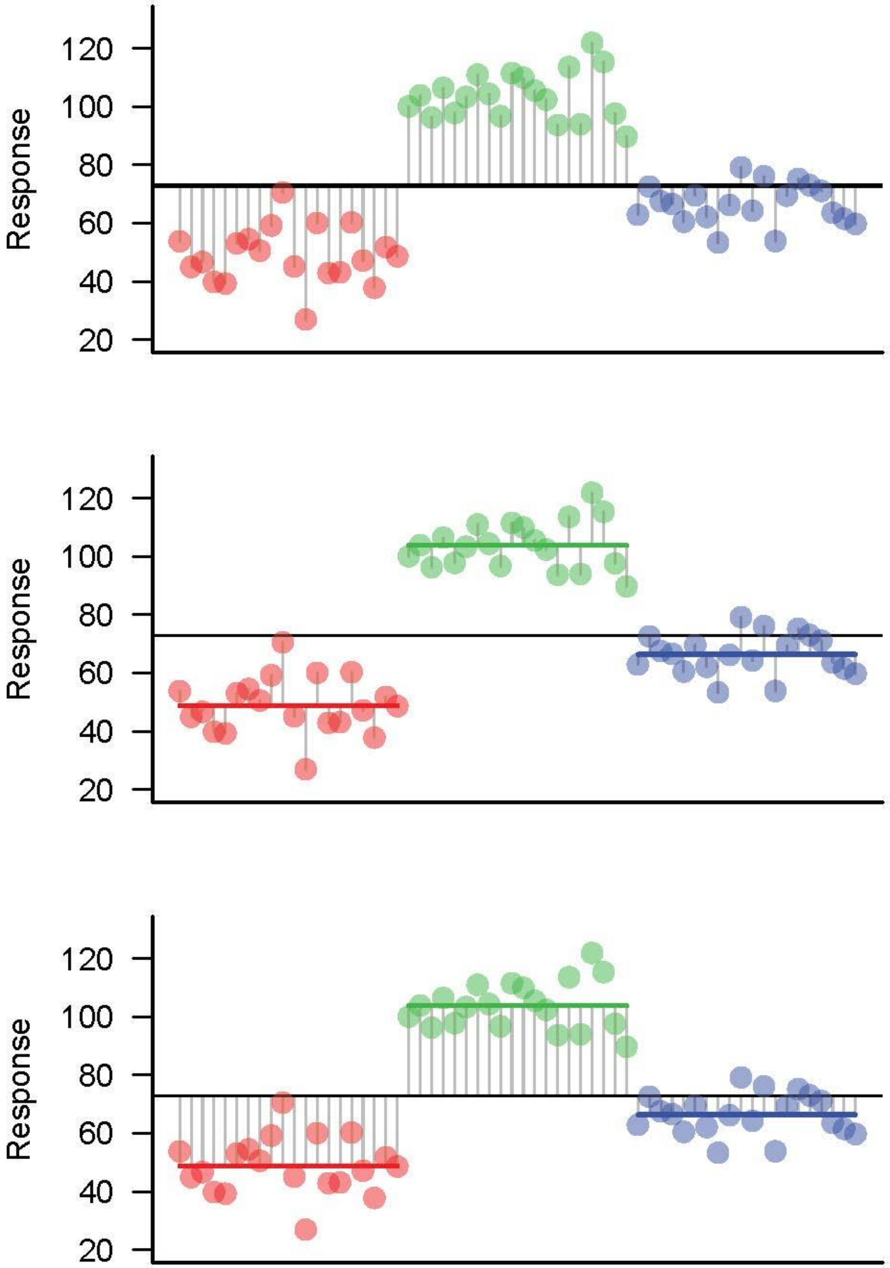
$$SS_W = \sum_g \sum_i \left(x_{ig} - \bar{x}_g\right)^2$$

- $df_W = g - 1$

$$SS_B = \sum_g n_g \left(\bar{x}_g - \bar{x}\right)^2$$

- $df_B = n - g$

g is the group; i is the observation



ANOVA – partitioning the variance

ANOVA uses the *F* statistic to determine if differences as significant:

$$F = \frac{MS_B}{MS_W}$$

Wait… what is *MS*?

The mean square (*MS*), which is:

$$MS = \frac{SS}{df}$$

ANOVA table

Source of variation	$\boldsymbol{SS}$	$\boldsymbol{df}$	$\boldsymbol{MS}$	$\boldsymbol{F}$	$\boldsymbol{p}$
Between	$\sum_g n_g\left(\bar{x}_g - \bar{x}\right)^2$	$n - g$	$\frac{SS_B}{df_B}$		
Within	$\sum_g \sum_i \left(x_{ig} - \bar{x}_g\right)^2$	$g - 1$	$\frac{SS_W}{df_W}$		
Total	$\sum (x - \bar{x})^2$	$n - 1$	$\frac{SS_T}{df_T}$		

ANOVA table

Source of variation	SS	df	MS	F	p
Between	$\sum_g n_g(\bar{x}_g - \bar{x})^2$	$n - g$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_W}$	
Within	$\sum_g \sum_i (x_{ig} - \bar{x}_g)^2$	$g - 1$	$\frac{SS_W}{df_W}$		
Total	$\sum (x - \bar{x})^2$	$n - 1$	$-$		

p is the probability of observing the *F* statistic with a given *df* if the null hypothesis is true

- H_0: there are no significant differences between means (i.e. all means equal)

ANOVA

H_0: there are no significant differences between all means (i.e. all means equal)
H_1: there are significant differences between means

When do we reject the null hypothesis?

- $F = \frac{MS_B}{MS_W}$ and when F is large, p is small
- If $p <$ alpha (e.g. 0.05) we *reject* the null hypothesis
- If $p >$ alpha (e.g. 0.05) we *fail to reject* the null hypothesis

But what does it mean if we reject the null hypothesis?

- We can't say *which* means are different!

ANOVA + Tukey HSD

Which means are different can be determined using a *post-hoc* test called:

Tukey's Honest Significant Difference test (Tukey HSD):

$$t = \frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{\frac{MS_W\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}{2}}}$$

- Accounts for the fact that multiple comparisons are being made (lower's possibility of type I errors)
- Calculates a t-statistic, BUT it treats each test as a pair of means, so $df = 1$

If $t > 4.303$, then $p < 0.05$

ANOVA?

But what if we have many things that could be affecting our response variable (several predictor variables)?

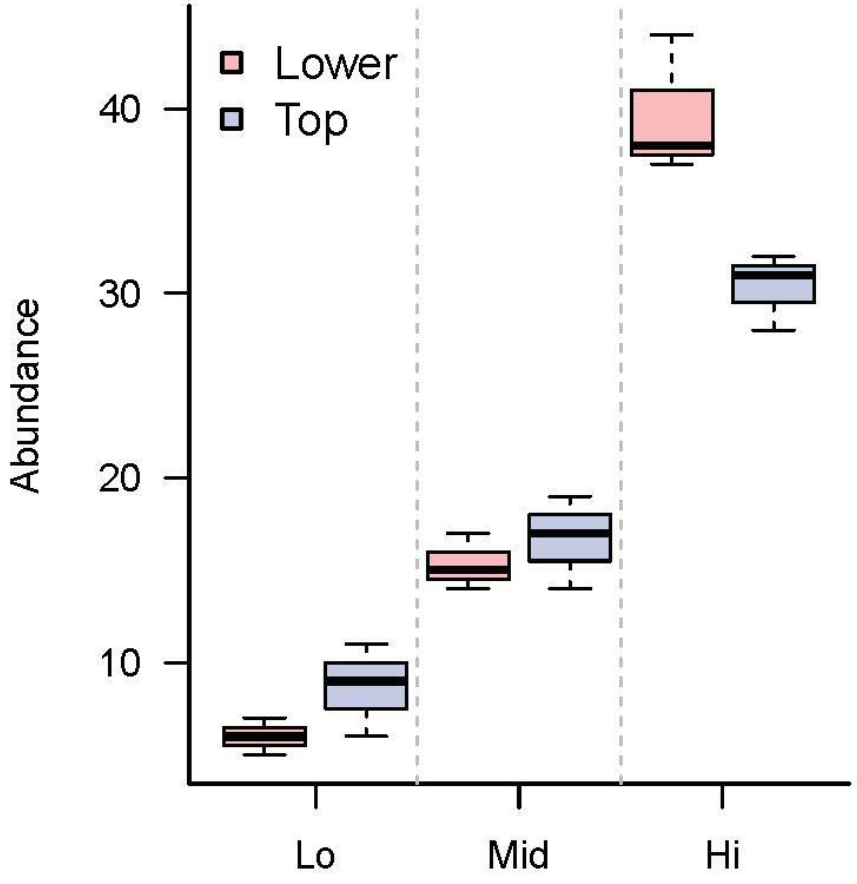
- Note: predictor variables that are categorical in ANOVAs = a "factor"
- 2 factors/predictor variables: *two-way* ANOVA
- 3 factors/predictor variables: *three-way* ANOVA
- … multi-way ANOVA

Two-way ANOVA example

We have two elevations where we sampled for invasive plant species abundance, lower elevation, and then at a top of a mountain ("Lower" and "Top" respectively). We also sampled at sites that had relatively low disturbance, medium level disturbance, and high levels of disturbance (i.e. "Lo", "Mid", "Hi").

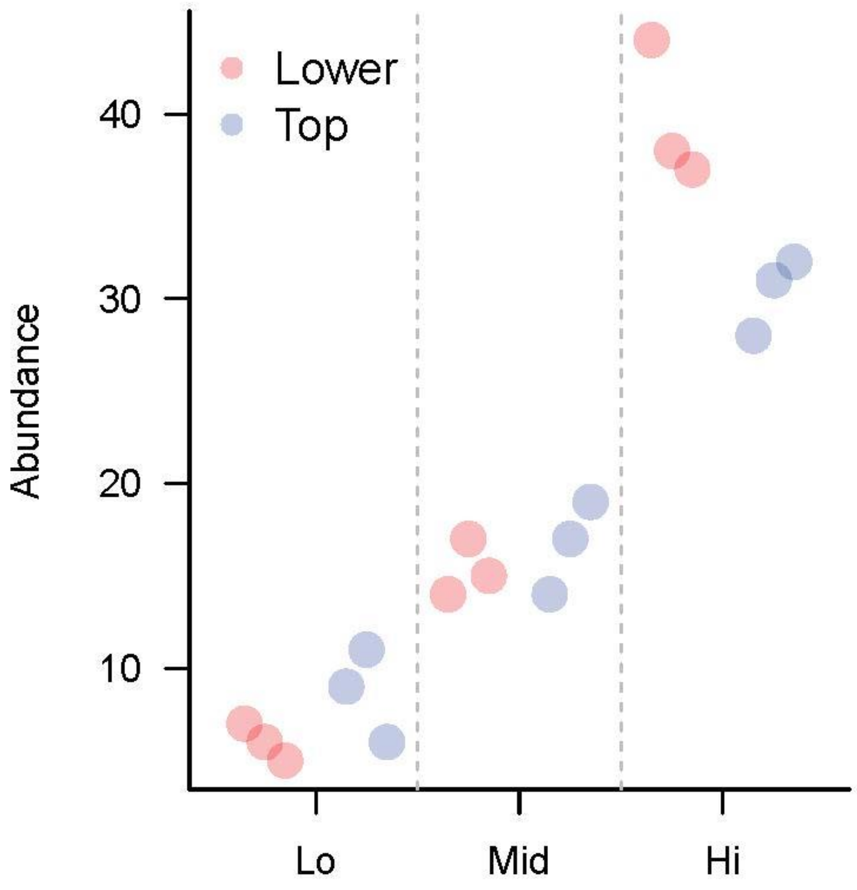
We want to answer the question:

Does disturbance or elevation have more of an effect on invasive species abundance, or does the interaction of the two matter?



Two-way ANOVA example

We have two elevations where we sampled for invasive plant species abundance, lower elevation, and then at a top of a mountain ("Lower" and "Top" respectively). We also sampled at sites that had relatively low disturbance, medium level disturbance, and high levels of disturbance (i.e. "Lo", "Mid", "Hi").



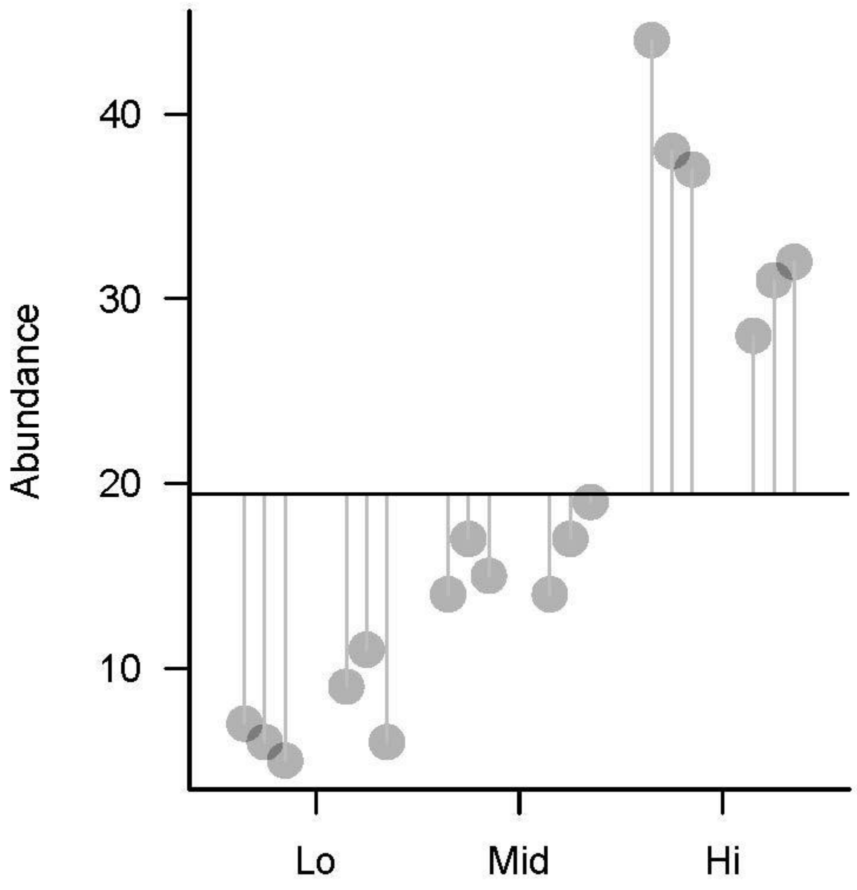
We want to answer the question:

Does disturbance or elevation have more of an effect on invasive species abundance, or does the interaction of the two matter?

Two-way ANOVA example

How do we partition the variance?

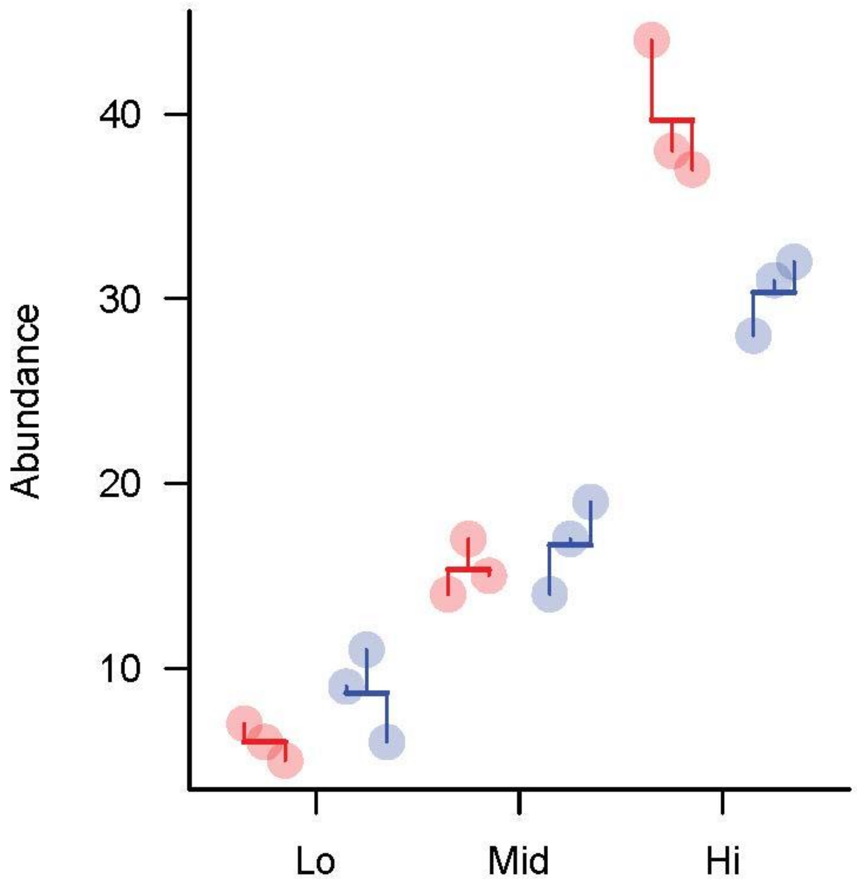
- SS total (like in the one-way)
  - $SS_T = \sum(x_i - \bar{x})^2$



Two-way ANOVA example

How do we partition the variance?

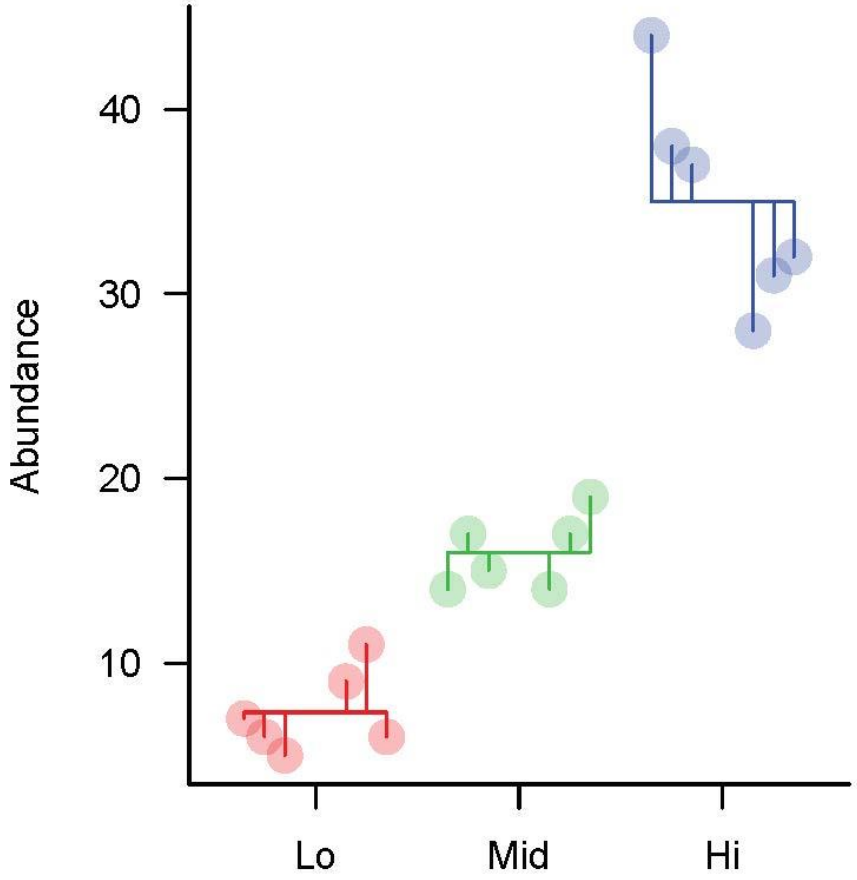
- SS total (like in the one-way)
  - $SS_T = \sum(x_i - \bar{x})^2$
- SS within all groups (like in the one-way)
  - $SS_W = \sum_g \sum_i \left(x_{ig} - \bar{x}_g\right)^2$



Two-way ANOVA example

How do we partition the variance?

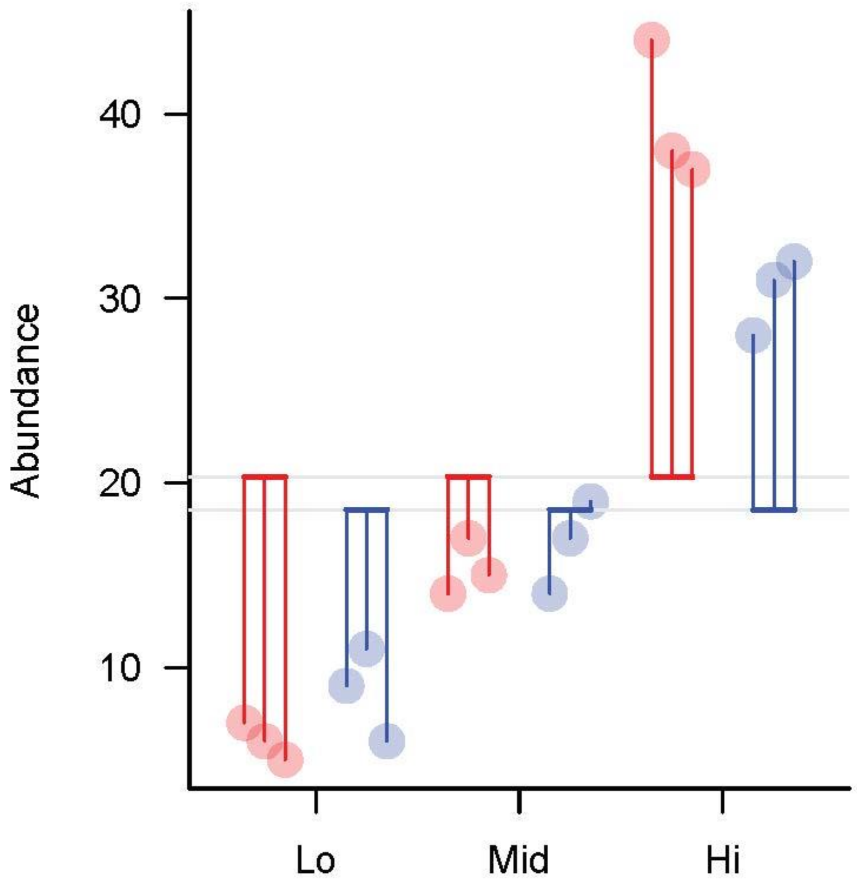
- SS total (like in the one-way)
  - $SS_T = \sum(x_i - \bar{x})^2$
- SS within all groups (like in the one-way)
  - $SS_W = \sum_g \sum_i \left(x_{ig} - \bar{x}_g\right)^2$
- SS for each factor
  - $SS_{disturb} = \sum\left(x_{i,disturb} - \bar{x}_{disturb}\right)^2$



Two-way ANOVA example

How do we partition the variance?

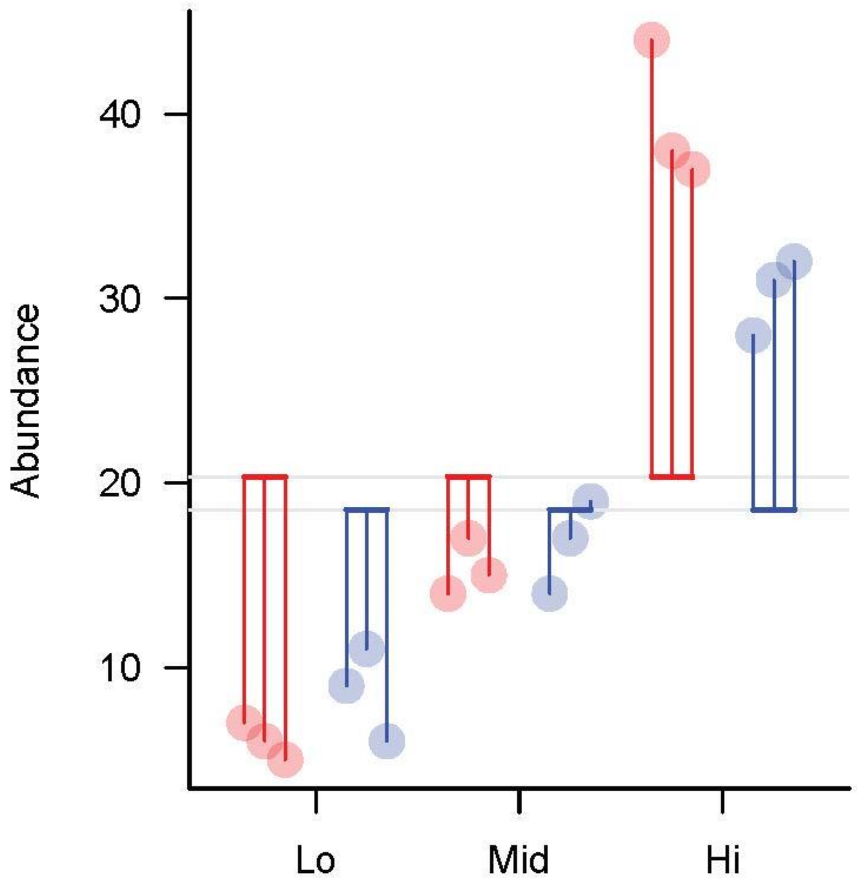
- SS total (like in the one-way)
  - $SS_T = \sum(x_i - \bar{x})^2$
- SS within all groups (like in the one-way)
  - $SS_W = \sum_g \sum_i \left(x_{ig} - \bar{x}_g\right)^2$
- SS for each factor
  - $SS_{disturb} = \sum\left(x_{i,disturb} - \bar{x}_{disturb}\right)^2$
  - $SS_{elevation} = \sum\left(x_{i,elevation} - \bar{x}_{elevation}\right)^2$



Two-way ANOVA example

How do we partition the variance?

- SS total (like in the one-way)
  - $SS_T = \sum(x_i - \bar{x})^2$
- SS within all groups (like in the one-way)
  - $SS_W = \sum_g \sum_i \left(x_{ig} - \bar{x}_g\right)^2$
- SS for each factor
  - $SS_{disturb} = \sum\left(x_{i,disturb} - \bar{x}_{disturb}\right)^2$
  - $SS_{elevation} = \sum\left(x_{i,elevation} - \bar{x}_{elevation}\right)^2$
- SS for the interaction of factors
  - $SS_{interaction} = SS_T - SS_W - SS_{disturb} - SS_{elevation}$



Two-way ANOVA table

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Disturbance	$\sum\left(x_{i,d} - \bar{x}_d\right)^2$	$l_d - 1$	$\frac{SS_d}{df_d}$	$\frac{MS_d}{MS_W}$	
Elevation	$\sum\left(x_{i,e} - \bar{x}_e\right)^2$	$l_e - 1$	$\frac{SS_e}{df_e}$	$\frac{MS_e}{MS_W}$	
Interaction	$SS_T - SS_W - SS_d - SS_e$	$df_d \times df_e$	$\frac{SS_i}{df_i}$	$\frac{MS_i}{MS_W}$	
Within	$\sum_g \sum_i \left(x_{ig} - \bar{x}_g\right)^2$	$n - (l_d \times l_e)$	$\frac{SS_W}{df_W}$		
Total	$\sum(x_i - \bar{x})^2$	$n - 1$			

Two-way ANOVA

- R for ANOVA:

```
#### two-way ANOVAs ----
## additive
twowayANOVA.additive <- aov(abundance ~ disturbance + elevation, data = inv)
summary(twowayANOVA.additive)

## interaction
twowayANOVA.interaction <- aov(abundance ~ disturbance * elevation, data = inv)
summary(twowayANOVA.interaction)
```

```
> summary(twowayANOVA.additive)
            Df Sum Sq Mean Sq F value   Pr(>F)    
disturbance  2 2403.1  1201.6   84.48 1.54e-08 ***
elevation    1   14.2    14.2    1.00    0.334    
Residuals   14  199.1    14.2                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(twowayANOVA.interaction)
                      Df Sum Sq Mean Sq F value   Pr(>F)    
disturbance            2 2403.1  1201.6 207.962 4.86e-10 ***
elevation              1   14.2    14.2   2.462  0.14264    
disturbance:elevation  2  129.8    64.9  11.231  0.00178 ** 
Residuals             12   69.3     5.8                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What happens when our data aren't normally distributed?

Kruskal-Wallis test!

Week 8 – Differences between more than two samples

Part II - Wednesday

Statistical testing

Using a t-test to test the difference between two samples, you calculate a *p*-value of 0.02. If you using a 5% significance (alpha) level – what do you conclude?

a) You reject the null hypothesis
b) You accept the null hypothesis
c) You don't reject the null hypothesis
d) You reject the alternative hypothesis
e) You don't reject the alternative hypothesis

iclicker®

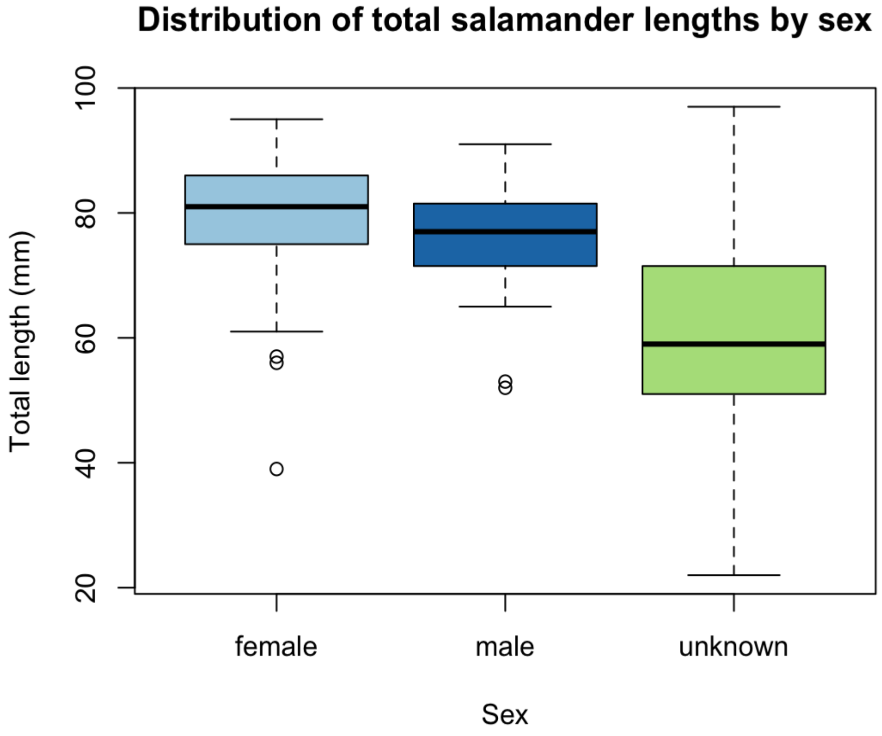
Today's Exercise

Back to the redbacked salamander (*Plethodon cinereus)* data! We're trying to answer the question:

Is the average length of salamanders significantly different between salamanders identified as male, female, or unknown and which are significantly different?

To do so:

1. Create a NEW R script (make sure you have your group names and date at the top!)!
   1. Do a one-way ANOVA and Tukey HSD test for the Total_length data
2. Write a <1 page report that includes the results from your ANOVA and Tukey HSD test and a short description of what they mean.
3. Upload your ***group*** code and report to moodle



For Monday:

1) Read Ch 8.2, 8.4 (skip 8.4.1), 8.5, 9.1 and 9.2 in Gardner (2017) – correlations and associations
2) Answer the individual evaluation questions on moodle

All before 11:55pm on Sunday