# Week 8: Differences among more than two samples

Session 1

Spring 2020

# iClicker Question 1

I want to compare fish weights in three lakes. I've sampled 32 fishes from each lake.

What statistical analysis should I perform?

- A `Multiple linear regression`
- B `t-test`
- C `one-way ANOVA`
- D `Chi-square test`

# iClicker Question 2

I want to compare fish weights in two lakes. I've sampled 32 fishes from each lake.

What statistical analysis should I perform?

- A `Multiple linear regression`
- B `t-test`
- C `one-way ANOVA`
- D `Chi-square test`

# Unit Topics

# Learning Objectives

# Comparing two groups

The t-test was limited to two groups, i.e. one factor with two levels. How could we test three levels using t-tests?

?

Analysis of Variance allows us to test multiple levels and factors simultaneously in a single statistical procedure.

# Comparing multiple groups - ANOVA

Analysis of Variance (ANOVA):

- A statistical test for testing for differences among 2 or more levels of a factor
- A statistical test for testing for differences in 1 or more factors
- ANOVA and t-test are identical when there is 1 factor with 2 levels.
- one factor/group/category (*One-way ANOVA*)
- two factor/group/category (*One-way ANOVA*)

The t-test was limited to two groups, i.e. one factor with two levels. How could we test three levels using t-tests?

Analysis of Variance allows us to test multiple levels and factors simultaneously in a single statistical procedure.

# ANOVA Definitions

- Factor: A **categorical** predictor variable
- Level: One possible value of a factor
- Group: A set of observations that share the same values for each factor
- *One-way ANOVA*: An ANOVA with a single factor
- *Two-way ANOVA*: An ANOVA with two factors
- Crossed design: An experiment with two or more factors with groups representing all possible combinations of factors/levels

# ANOVA Assumptions

The basic Analysis of Variance model includes the following assumptions:

- The data are **normally distributed** within groups.
- The data have the same **standard deviation** within groups.
- All observations are **independent**.
- All groups have the same number of samples.

There are variations on the standard ANOVA to accommodate violations of each of these assumptions.

# ANOVA Hypotheses

We can think of ANOVA as an extension of the t-test.

What was the t-test **null** hypothesis?

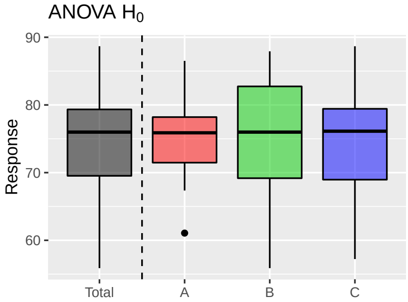
What were the possible t-test **alternative** hypotheses?

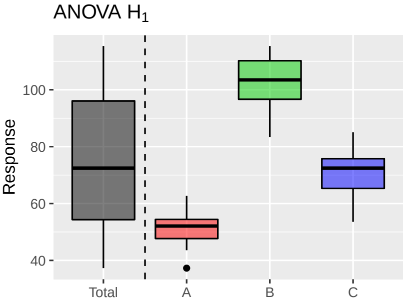
Keeping these hypotheses in mind, how could we define null and alternative hypotheses for an ANOVA?

?

# ANOVA Hypotheses

- $H_0$: There are no significant differences among groups.
  - All group means are equal.
  - All **samples** come from the **same population**.
- $H_1$: There are significant differences among the groups.
  - All group means are not equal.
  - Some **samples** come from **different populations**.



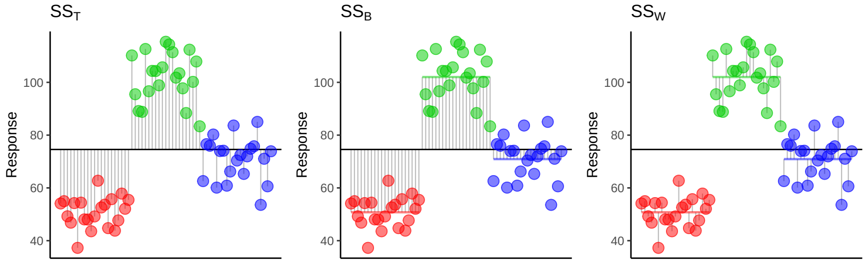



# ANOVA explained

So what is an ANOVA actually doing?

In any statistical model, we want to explain the **variation** we observe in our data.

The ANOVA partitions the **total** variation into **between** sample (group) variation with **within** sample (group) variation to determine whether samples come from a single distribution or not.

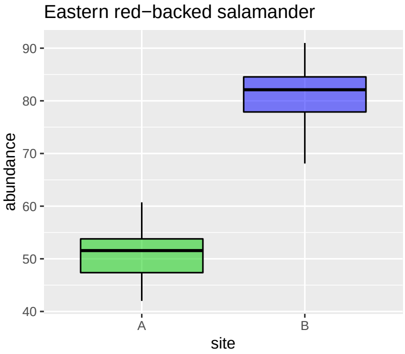


# Comparing differences

Two samples: Salamander abundance at two sites

- Which test do we use?

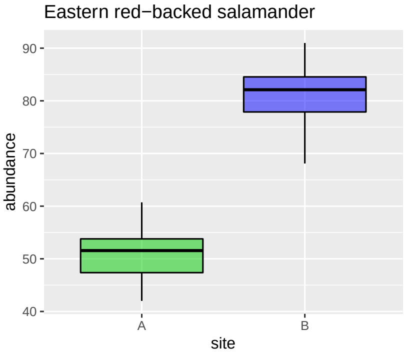
?



# Comparing differences - two samples

Two samples:

- the t-test?
- test whether group means differ significantly
- $H_0$: there is no significant difference between the means
- $H_1$: there is a significant difference between the means



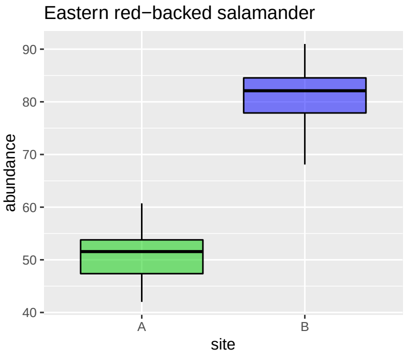
# Comparing differences - two smaples

Two samples:

- the t-test?
- test whether group means differ significantly
- $H_0$: there is no significant difference between the means
- $H_1$: there is a significant difference between the means

Significance based on:

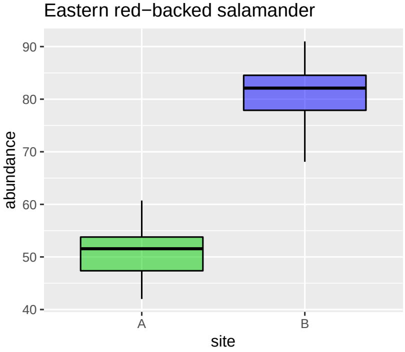
- t-statistic: $t = \frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$
- degrees of freedom
- $p$-value



# Comparing differences - more than two samples

What about if there are more than 2 samples?

- Can you think of any examples?



# Moving beyond two groups

T-tests are great, but what if we need to analyze more complicated scenarios?

Let's walk through some sampling and experimental scenarios to build intuition.

Scenario context: We're interested in bluegill population densities in Massachusetts lakes.


[1] Image credit: New York Fish and Game Commission

# Scenario 1

Having just analyzed some fish counts data in 16 lakes in Massachusetts, Thorsten found a significant 'lake' effect using an ANOVA, i.e., the mean number of fish was not the same in all lakes.

1. Thorsten wants to know which *which lakes are different from each other*.
   - Think carefully: what does this actually mean?
   - What is the sampling unit?
   - What did he measure?
   - What would he need to compare?

# Scenario 1

Having just analyzed some fish counts data in 16 lakes in Massachusetts, Thorsten found a significant 'lake' effect using an ANOVA, i.e., the mean number of fish was not the same in all lakes.

2. What would Thorsten do to find out *which lakes were different from eachother*?

A) A series of $t$-tests

B) A Tukey Honest Significant Difference test

C) A Kruskal-Wallis test

## Scenario 2

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:

1. What is different from the last scenario?
2. What is the sampling unit?
3. What specific question(s) should I ask?
4. What, specifically, do I want to compare?

# Scenario 2

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:

5. Which statistical test could I use?

A) A $t$-test
B) A One-Way ANOVA
C) A Chi-square test
D) A Two-Way ANOVA

6. Which is the test statistic for the test I chose?

# Scenario 3

I am interested in testing whether there is a significant difference between the population density 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**.

1. How has our sampling scheme changed?
2. What is the sampling unit?
3. How has our question changed?

# Scenario 3

I am interested in testing whether there is a significant difference between the population density 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**.

4. Now which statistical test should I use?

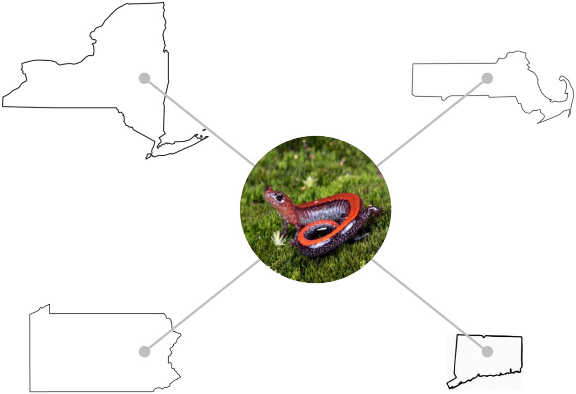
A) A $t$-test
B) A One-Way ANOVA
C) A Chi-square test
D) A Two-Way ANOVA

5. Now which is the test statistic for the test?

# Comparing multiple groups - examples

Regional differences in salamander abundance:

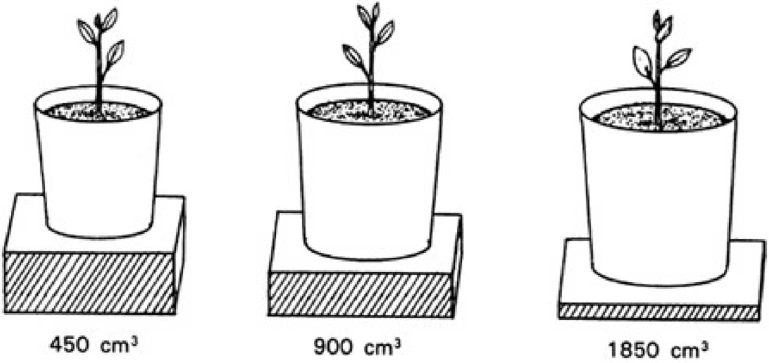
- comparing multiple populations
- quantify the differences between populations



# Comparing multiple groups - examples

Plant growth related to available resources (pot size):

- comparing multiple treatments
- quantify the effects of resource availability



# Comparing multiple groups - examples

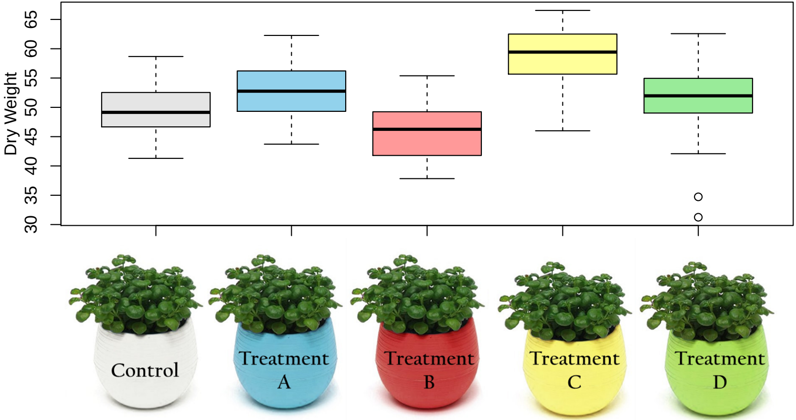
Plants productivity (dry mass in grams) related to fertilizer treatment

- do our treatments influence biomass production?
- is there a positive effect relative to a control?

# Comparing multiple groups - examples

When there are more than 2 groups

- t-test is probably not optimal:
- We would need to do all possible pairs.
- We might get spurious differences just by chance. Why?



# Set notation: review

Example

Say we have a collection of 5 numbers: $X$ (uppercase).

$$X = \{1, 3, 5, 4, 6\}$$

- We'll define $n$ to be the count of numbers: $n = 5$.
- We can call $X$ a **set** with 5 elements. - Note the curly braces and uppercase $X$ to indicate the whole set. - We use a lowercase $x$ and a subscript $i$ to indicate a specific **element** in set $X$: $x_i$

$$x_1 = 1, x_2 = 3, x_3 = 5, ...$$

# Sigma notation: review

We use the capital greek letter **Sigma** as a shorthand for the **sum** of a series of numbers:

$$\sum X = \sum_{i}^{n} x = 19$$

- Note the uppercase $X$, lowercase $x$, and index subscript $i$.

We can also use a **function** with **sigma notation** as a shorthand for the sum of the **function** evaluated for each element in $X$. For example, the sum of squared elements in $X$ would be:

$$\sum_{i}^{n} x^2 = 1^2 + 3^2 + 5^2 + 4^2 + 6^2 = 87$$

# Sums of Squares: Intuition

We often use **Sums of Squared Differences** to measure **variability** in a set of observations. We're usually interested in the differences from a **mean**:

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

Notable properties of Sums of Squares

- SS are small when observations are close to the mean.
- SS are large when observations are far from the mean.
- Because of the **squaring**, observations far from the mean make large contributions to the sum.

# Sums of Squares: Intuition

Sums of Squares can help us estimate our **model goodness-of-fit**.

Suppose we have a collection of $i$ observations: $x_1, x_2, ...x_i$
What quantity of $y$ will **minimize** the following sum?

$$\sum_{i}^{n}(x_i - y)^2$$

?

# Sums of Squares: Intuition

It turns out that the mean value of all the $x_i$ results in the minimum possible $SS$!

$$min(SS) = \sum_{i}^{n}(x_i - \bar{x})^2$$

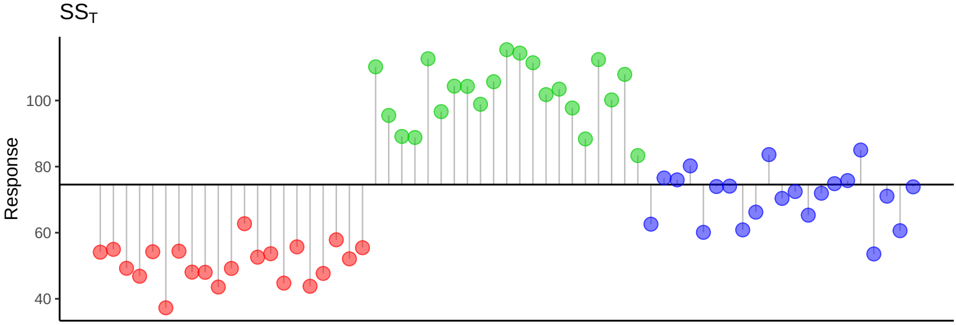
The **mean** is the *best* description of the central tendency because it **minimizes** the $SS$.

We can try to minimize the Sums of Squares in more complicated models too.

# ANOVA: Total Sums of Squares ($SS_T$)

- $N$ is the total number of observations, $\bar{x}$ is the **grand mean**

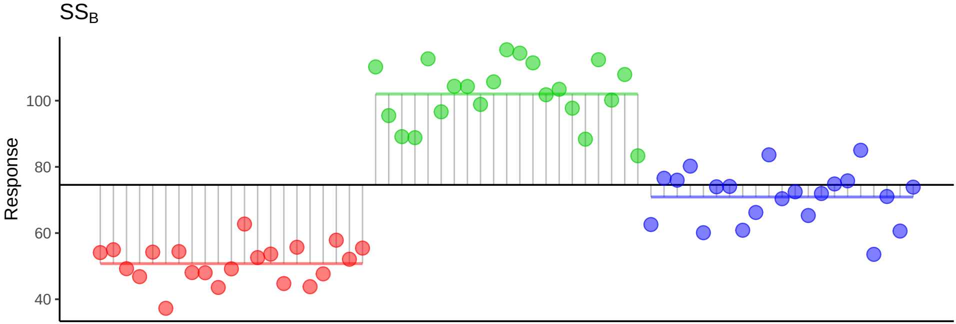
$$SS_T = \sum_{i}^{N} (x_i - \bar{x})^2$$



# ANOVA: Between-sample Sums of Squares ($SS_B$)

- Add up the differences between the **group means** and the **grand mean**.

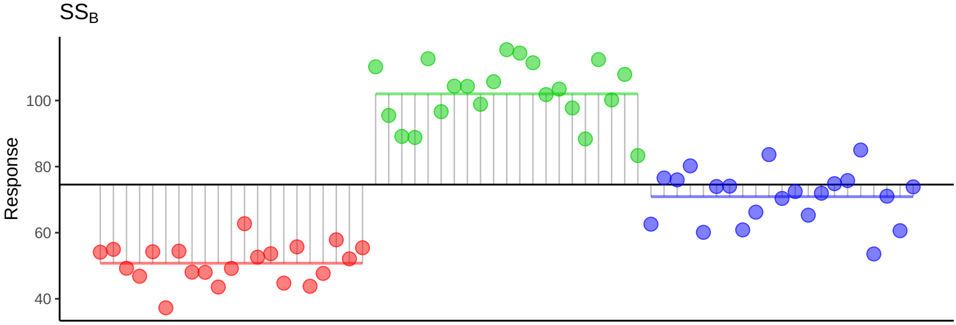
$$SS_B = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2$$



# ANOVA: Between-sample Sums of Squares ($SS_B$)

- $g$ is the number of groups, $n_j$ is the number of observations in group $j$

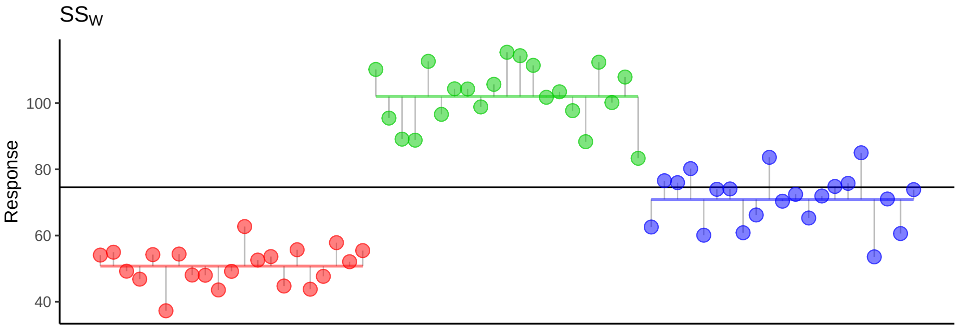
$$SS_B = \sum_{j}^{g} n_j(\bar{x}_j - \bar{x})^2$$



# ANOVA: Within-sample Sums of Squares ($SS_W$)

- Calculate the **SS** separately for each group using the **group mean**.

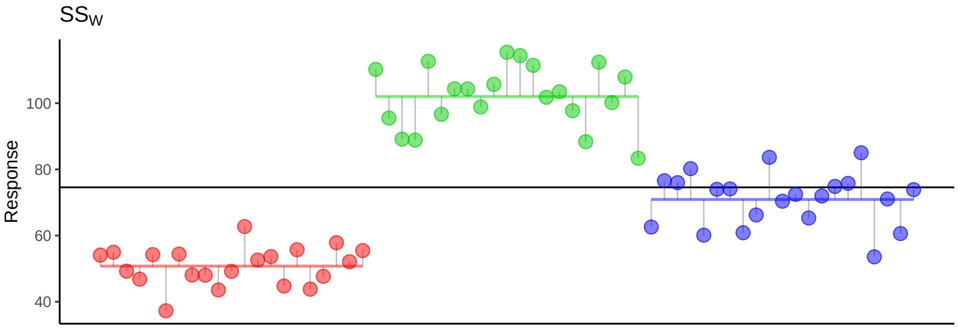
$$SS_W = \sum_{i}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i}^{n_2}(x_{i2} - \bar{x}_2)^2 + \sum_{i}^{n_3}(x_{i3} - \bar{x}_3)^2$$



# ANOVA: Within-sample Sums of Squares ($SS_W$)

- $g$ is the number of groups, $n_j$ is the number of observations in group $j$

$$SS_W = \sum_{j}^{g} \sum_{i}^{n_j} (x_{ij} - \bar{x}_j)^2$$



# Sums of Squares terminology

Confusingly, the different terms are sometimes used for the different $SS$ terms:

- The between-group $SS$ may also be called the **factor** or **main effect** sum of squares.
- The within-group $SS$ is also known as the **error** sum of squares and the **residual** sum of squares.

# ANOVA: Minimizing the Sum of Squares

Remember minimizing the Sum of Squares to find the mean of a set of numbers?

# ANOVA degrees of freedom

If we define the following:

- $N$ (uppercase) is the total sample size (number of observations)
- $g$ is the number of groups/samples
- $n_j$ (lowercase) is the number of samples in group $j$.

Then the degrees of freedom ($df$) are:

- Total: $df_T = N - 1$
- Within: $df_W = g - 1$
- Between: $df_B = n - g$ - when all groups have the same number of samples ($n$).
- Residuals: $df_r = df_T - df_B$

NOTE: We haven't talked about **residuals** yet.

# Deegrees of freedom

But what are **degrees of freedom**?

?

# Deegrees of freedom

The statistical theoretical concept of degrees of freedom can be difficult to grasp. You can just think of them as an adjusted version of the sample size.

When we estimate **popoulation parameters** from **sample statistics**, we want to avoid **bias**. The smaller the sample size, the more likely to have bias.

Using degrees of freedom, rather than sample sizes helps to compensate for bias.

# ANOVA the *mean square*

The mean square ($MS$) is the sum of squares divided by the degrees of freedom:

$$MS = SS/df$$

So:

- Total: $MS_T = SS_T/df_T$
- Within: $MS_W = SS_W/df_W$
- Between: $MS_B = SS_B/df_B$

Dividing by the degrees of freedom helps to **normalize** the sums of squares so that we can compare pairs of $SS$'s, such as $SS_W$ and $SS_B$, directly.

# ANOVA all the ingredients

| | $SS$ | $df$ | $MS$ |
|---|---|---|---|
| Total | $\sum(x-\bar{x})^2$ | $n-1$ | $SS_T/df_T$ |
| Within | $\sum_g \sum_i (x_{ig}-\bar{x}_j)^2$ | $n-g$ | $SS_W/df_W$ |
| Between | $\sum_g n_g(\bar{x}_g-\bar{x})^2$ | $g-1$ | $SS_B/df_B$ |

# ANOVA the statistical test

ANOVA results are usually presented in an ANOVA table

| Source of variation | $SS$ | $df$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| Between | $SS_B$ | $df_B$ | $MS_B$ | | |
| Within | $SS_W$ | $df_W$ | $MS_W$ | | |
| Total | $SS_T$ | $df_T$ | – | | |

# ANOVA the statistical test

ANOVA results are usually presented in an ANOVA table

| Source of variation | $SS$ | $df$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| Between | $SS_B$ | $df_B$ | $MS_B$ | | |
| Within | $SS_W$ | $df_W$ | $MS_W$ | | |
| Total | $SS_T$ | $df_T$ | – | | |

- $F$ is the test statistic for the ANOVA

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

# ANOVA the statistical test

ANOVA results are usually presented in an ANOVA table

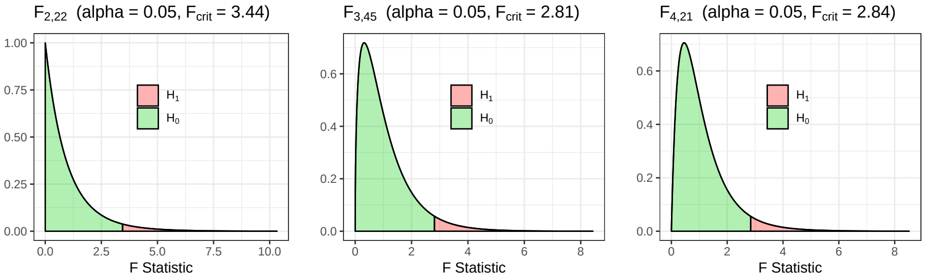
| Source of variation | $SS$ | $df$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| Between | $SS_B$ | $df_B$ | $MS_B$ | | |
| Within | $SS_W$ | $df_W$ | $MS_W$ | | |
| Total | $SS_T$ | $df_T$ | – | | |

- $p$ is the probability of observing the $F$ statistic with a given degrees of freedom if the null hypothesis is true:
  - null hypothesis is 'no difference between the means'
  - based on the $F$-distribution

# ANOVA the $F$ distribution

The $F$ distribution is actually a family of distributions, parameterized by the **numerator degress of freedom** and the **denominator degrees of freedom**.

An $F$ statistic is just a ratio of two **mean squares** terms, hence the **numerator** and **denominator** degrees of freedom.

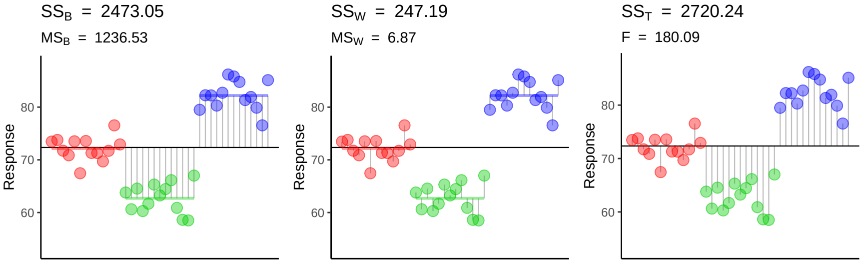


# ANOVA and the Sums of Squares

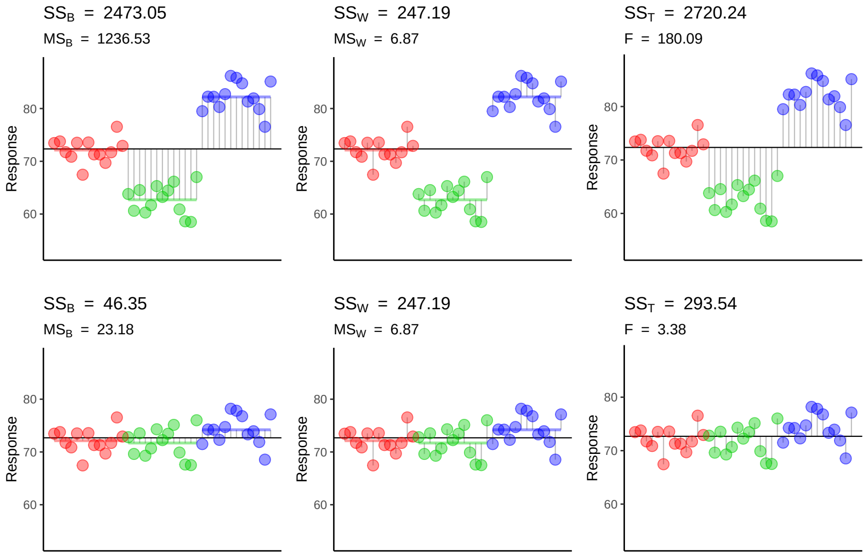
Analysis of Variance **partitions** the total variation ($SS_T$) into within- and between-group variation.

The $F$ statistic is the ratio!

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



# ANOVA and the Sums of Squares



# ANOVA the $p$ value

Hypotheses:

- $H_0$: there are no significant differences between the means
  - all means are equal
- $H_1$: there are significant differences between the means
  - all means are not equal

When do we reject or fail to reject the null hypothesis?

# ANOVA the $p$ value

Hypotheses:

- $H_0$: there are no significant differences between the means
  - all means are equal
- $H_1$: there are significant differences between the means
  - all means are not equal

When do we reject or fail to reject the null hypothesis?

- if $F$ is large, then $p$ is small
- if $p < 0.05$ we reject the null hypothesis
- if $p > 0.05$ we *fail to* reject the null hypothesis

# Pairwise comparisons with ANOVA

The $F$ statistic tells us whether there are differences, but *not* what the differences are:

- *Cannot* use $t$-tests to make pairwise comparisons
  - multiple $t$-tests will lead to significant results by chance

# Pairwise comparisons with ANOVA

The $F$ statistic tells us whether there are differences, but *not* what the differences are:

- Instead we conduct *Post-hoc* testing
  - Tukey Honest Significant Difference test (Tukey HSD)
  - accounts for multiple tests being conducted
  - calculation of a $t$-statistic
  - a pair, so degrees of freedom is 1
  - 5% critical value for $df = 1$ is 4.303
    - if $t > 4.303$ then $p < 0.05$

# Pairwise comparisons with ANOVA

The $F$ statistic tells us whether there are differences, but *not* what the differences are:

- Instead we conduct *Post-hoc* testing
  - Tukey Honest Significant Difference test (Tukey HSD)
  - accounts for multiple tests being conducted
  - calculation of a $t$-statistic
  - a pair, so degrees of freedom is 1
  - 5% critical value for $df = 1$ is 4.303
    - if $t > 4.303$ then $p < 0.05$

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

# Pairwise comparisons with ANOVA

The $F$ statistic tells us whether there are differences, but *not* what the differences are:

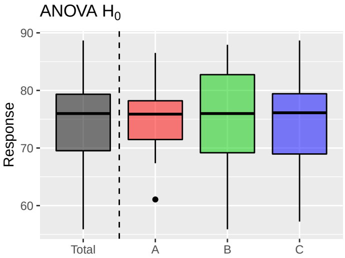
- Instead we conduct *Post-hoc* testing
  - Tukey Honest Significant Difference test (Tukey HSD)
  - accounts for multiple tests being conducted
  - calculation of a $t$-statistic
  - a pair, so degrees of freedom is 1
  - 5% critical value for $df = 1$ is 4.303
    - if $t > 4.303$ then $p < 0.05$

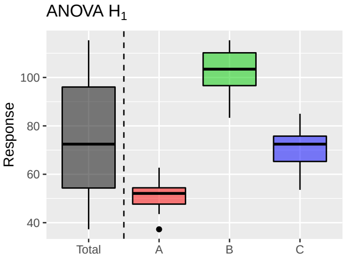
| | A | B | C |
|---|---|---|---|
| A | - | $t_{A,B}$ | $t_{A,C}$ |
| B | - | - | $t_{B,C}$ |
| C | - | - | - |

# ANOVA Recap

Comparing differences between >2 samples (groups) using ANOVA

- null hypothesis:
  - no difference between the samples
  - data are from the same population
- alternative hypothesis:
  - sample means are different
  - data from the different populations

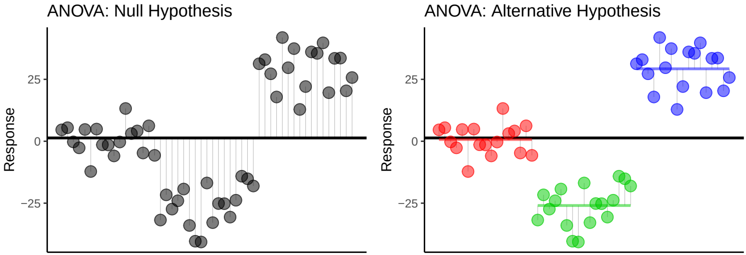


# ANOVA Recap

Comparing differences between >2 groups using ANOVA

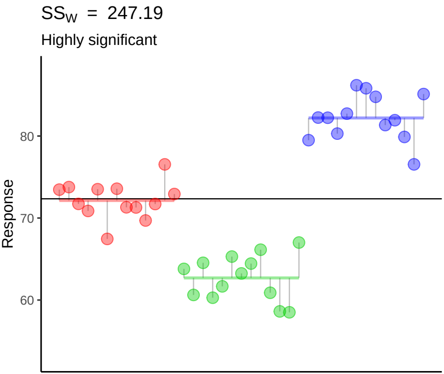
| Source of variation | $SS$ | $df$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| Between | $SS_B$ | $df_B$ | $MS_B$ | $\frac{MS_B}{MS_W}$ | |
| Within | $SS_W$ | $df_W$ | $MS_W$ | | |
| Total | $SS_T$ | $df_T$ | – | | |

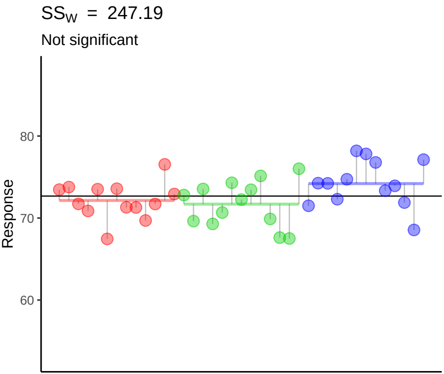


# ANOVA Recap

Comparing differences between >2 groups using ANOVA

- Essentially comes down to:
  - a model with one mean *or* a model with a mean per group
  - which model best explains the data
  - which model significantly reduces the sums of squares



# More than one factor with ANOVA

So far we have looked at multiple levels within a single factor

- factor: a single categorical predictor variable
- level: the categories within a factor

In some cases, we may be interested in >1 factor

- 2 factors: *two-way* ANOVA
- 3 factors: *three-way* ANOVA
- $\cdots$ multi-way ANOVA

# Two-way ANOVA

Let's use a grazing example:

| | Site | |
|---|---|---|
| Grazing Treatment | Top | Lower |
| Lo | 9 | 7 |
| Lo | 11 | 6 |
| Lo | 6 | 5 |
| Mid | 14 | 14 |
| Mid | 17 | 17 |
| Mid | 19 | 15 |
| Hi | 28 | 44 |
| Hi | 31 | 38 |
| Hi | 32 | 37 |

# Two-way ANOVA

Lets use the example from the book (in R looks like this):

```
graze
```

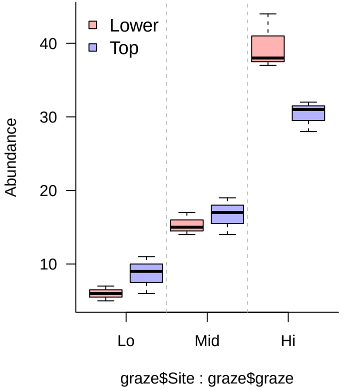
```
##    graze  Site Abundance
## 1     Lo   Top         9
## 2     Lo   Top        11
## 3     Lo   Top         6
## 4    Mid   Top        14
## 5    Mid   Top        17
## 6    Mid   Top        19
## 7     Hi   Top        28
## 8     Hi   Top        31
## 9     Hi   Top        32
## 10    Lo Lower         7
## 11    Lo Lower         6
## 12    Lo Lower         5
## 13   Mid Lower        14
```

# Two-way ANOVA

Lets use the example from the book (in R looks like this):

```
graze
```

```
##    graze  Site Abundance
## 1     Lo   Top         9
## 2     Lo   Top        11
## 3     Lo   Top         6
## 4    Mid   Top        14
## 5    Mid   Top        17
## 6    Mid   Top        19
## 7     Hi   Top        28
## 8     Hi   Top        31
## 9     Hi   Top        32
## 10    Lo Lower         7
## 11    Lo Lower         6
## 12    Lo Lower         5
## 13   Mid Lower        14
```

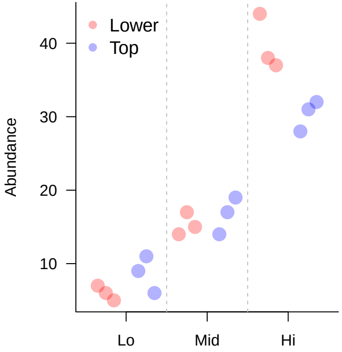


# Two-way ANOVA

Lets use the example from the book (in R looks like this):

```
head(graze, 9)
```

```
##   graze Site Abundance
## 1    Lo  Top         9
## 2    Lo  Top        11
## 3    Lo  Top         6
## 4   Mid  Top        14
## 5   Mid  Top        17
## 6   Mid  Top        19
## 7    Hi  Top        28
## 8    Hi  Top        31
## 9    Hi  Top        32
```

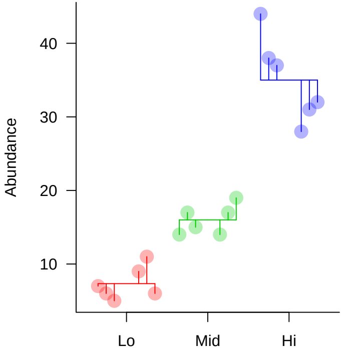



S

# Conducting the ANOVA

Step one:

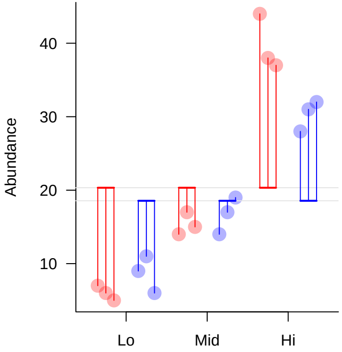
- SS for each factor
  - graze
  - site
- $SS_{graze} = \sum(x_{i,graze} - \bar{x}_{graze})^2$
- Ignore site grouping



# Conducting the ANOVA

Step one:

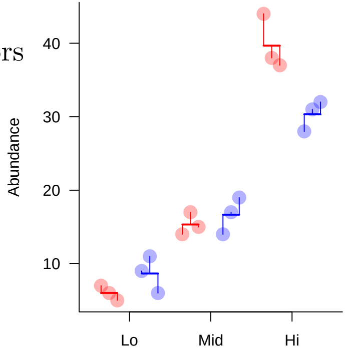
- SS for each factor
  - graze
  - site
- $SS_{site} = \sum(x_{i,site} - \bar{x}_{site})^2$
- Ignore graze grouping



# Conducting the ANOVA

Step two:

- SS for each combinations of factors
- Treat all groupings as unique
- $SS_{within} = (x_{i,g} - \bar{x}_g)^2$



# Conducting the ANOVA

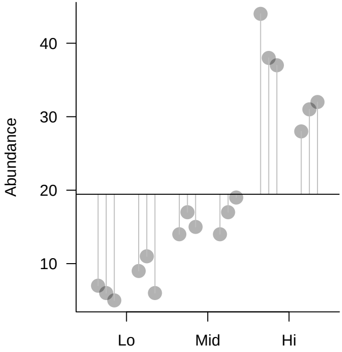
Step three:

- Sums of squares of both factors
- $SS_{both} = SS_{total} - SS_{graze} - SSsite - SS_{within}$

# Conducting the ANOVA

Step four:

- Total sums of squares
- $SS_{total} = \sum(x_i - \bar{x})^2$
- the *null* model
- Ignore all group structure



# Conducting the ANOVA - sums of squares

| | $SS$ | $df$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| Graze | $SS_{graze}$ | | | | |
| Site | $SS_{site}$ | | | | |
| Both factors(interaction) | $SS_{both}$ | | | | |
| Within group | $SS_{within}$ | | | | |
| Total | $SS_{total}$ | | | | |

# Degrees of freedom

In general:

- Factor 1 (F1): number of levels - 1
- Factor 2 (F2): number of levels - 1
- Within: $n$ - (levels in F1 $\times$ levels in F2)
- Total: $n$ - 1

# Degrees of freedom

In general:

- Factor 1 (F1): number of levels - 1
- Factor 2 (F2): number of levels - 1
- Within: $n$ - (levels in F1 $\times$ levels in F2)
- Total: $n$ - 1

Grazing example:

- Graze: $3 - 1 = 2$
- Site: $2 - 1 = 1$
- Within: $18 - (3 \times 2) = 12$
- Total: $18 - 1 = 17$

# Degrees of freedom

In general:

- Factor 1 (F1): number of levels - 1
- Factor 2 (F2): number of levels - 1
- Within: $n$ - (levels in F1 $\times$ levels in F2)
- Total: $n$ - 1

| | $SS$ | $df$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| Graze | $SS_{graze}$ | $df_{graze}$ | | | |
| Site | $SS_{site}$ | $df_{site}$ | | | |
| Both factors(interaction) | $SS_{both}$ | $df_{both}$ | | | |
| Within group | $SS_{within}$ | $df_{within}$ | | | |
| Total | $SS_{total}$ | $df_{total}$ | | | |

# Mean squares

- the mean squares are calculated by dividing the sums of squares by the degrees of freedom for each element

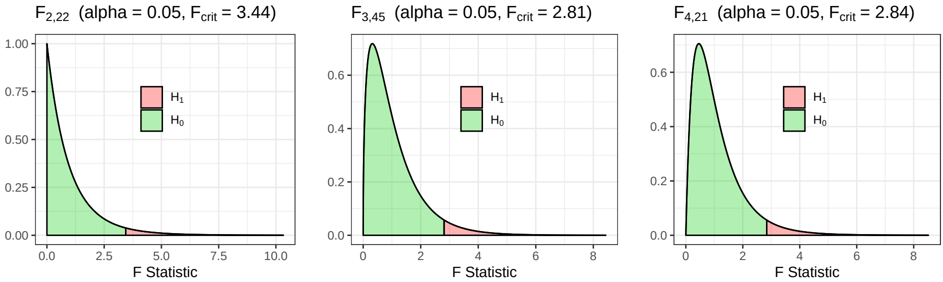
| | $SS$ | $df$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| Graze | $SS_{graze}$ | $df_{graze}$ | $MS_{graze} = \frac{SS_{graze}}{df_{graze}}$ | | |
| Site | $SS_{site}$ | $df_{site}$ | $MS_{site} = \frac{SS_{site}}{df_{site}}$ | | |
| Both factors | $SS_{both}$ | $df_{both}$ | $MS_{both} = \frac{SS_{both}}{df_{both}}$ | | |
| Within group | $SS_{within}$ | $df_{within}$ | $MS_{within} = \frac{SS_{within}}{df_{within}}$ | | |
| Total | $SS_{total}$ | $df_{total}$ | | | |

# F statistic

- the $F$-statistic is calculated by taking the element of interest divided by the within group MS (the *error* term)

| | $SS$ | $df$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| Graze | $SS_{graze}$ | $df_{graze}$ | $MS_{graze} = \frac{SS_{graze}}{df_{graze}}$ | $\frac{MS_{graze}}{MS_{within}}$ | |
| Site | $SS_{site}$ | $df_{site}$ | $MS_{site} = \frac{SS_{site}}{df_{site}}$ | $\frac{MS_{site}}{MS_{within}}$ | |
| Both factors | $SS_{both}$ | $df_{both}$ | $MS_{both} = \frac{SS_{both}}{df_{both}}$ | $\frac{MS_{both}}{MS_{within}}$ | |
| Within group | $SS_{within}$ | $df_{within}$ | $MS_{within} = \frac{SS_{within}}{df_{within}}$ | | |
| Total | $SS_{total}$ | $df_{total}$ | | | |

# ANOVA the $F$ distribution



# ANOVA in practice - R

- Read in the data as a data frame

```
head(graze, 10)
```

```
##    graze  Site Abundance
## 1     Lo   Top         9
## 2     Lo   Top        11
## 3     Lo   Top         6
## 4    Mid   Top        14
## 5    Mid   Top        17
## 6    Mid   Top        19
## 7     Hi   Top        28
## 8     Hi   Top        31
## 9     Hi   Top        32
## 10    Lo Lower         7
```

# Any ANOVA in practice - R

- Conduct *any 1-way* test using formula syntax

```r
oneway_fit = lm(Abundance ~ Site, data = graze)
anova(oneway_fit)
```

```
## Analysis of Variance Table
##
## Response: Abundance
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Site       1   14.22  14.222  0.0874 0.7713
## Residuals 16 2602.22 162.639
```

# 2-way Additive ANOVA in R

- Conduct a *2-way* additive test (without interaction) using formula syntax

```
twoway_additive = lm(aov(Abundance ~ Site + graze, data = g
anova(twoway_additive)
```

```
## Analysis of Variance Table
##
## Response: Abundance
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Site       1   14.22   14.22   1.000    0.3343
## graze      2 2403.11 1201.56  84.484 1.536e-08 ***
## Residuals 14  199.11   14.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

# 2-way Interaction ANOVA in R

- Conduct *2-way* test with **interaction** using formula syntax

```
twoway_interaction = lm(Abundance ~ Site * graze, data = gr
anova(twoway_interaction)
```

```
## Analysis of Variance Table
##
## Response: Abundance
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## Site         1   14.22   14.22   2.4615  0.142644
## graze        2 2403.11 1201.56 207.9615 4.863e-10 ***
## Site:graze   2  129.78   64.89  11.2308  0.001783 **
## Residuals   12   69.33    5.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

NOTE: to tell R to use an interaction, you use the star symbol rather than the + symbol.