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`

Announcements

Update to week 8 assignment

This week's material builds on ideas from pairwise group comparisons.

The ANOVA material is more *dense* than what we've covered up until this point.

We're going to have to work on our *statistical intuition* to master these *inferential statistics* concepts.

This week

Tuesday: Differences between more than two samples:

- Analysis of Variance (ANOVA) concepts
  - One-way ANOVA
  - Two-way ANOVA
  - Multipe testing

Thursday

- Continue ANOVA concepts
- Statistical analysis of salamanders

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?

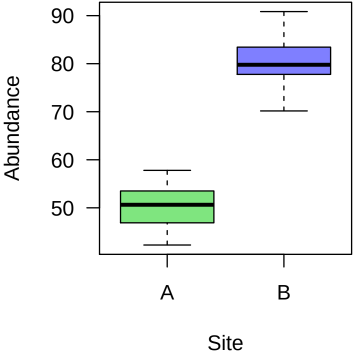
A) A t-test
B) A One-Way ANOVA
C) A Chi-square test
D) A Two-Way ANOVA

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

Comparing differences - two samples

Two samples:

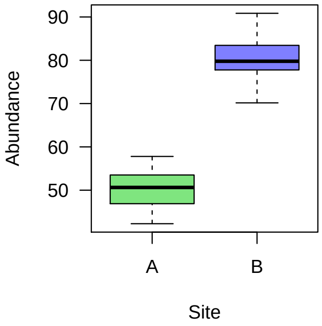
- Which test do we use?



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



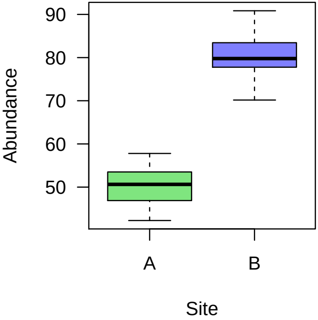
Comparing differences - two smaples

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?

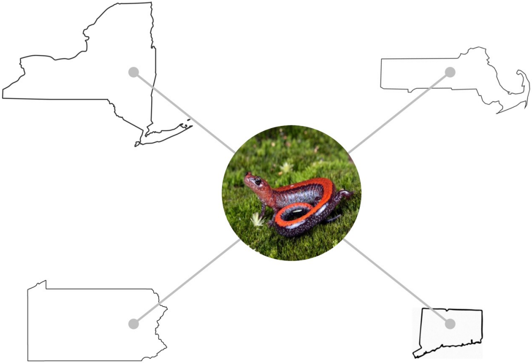
- can you think of any examples?

?

Comparing multiple groups - examples

Regional differences in salamander abundance:

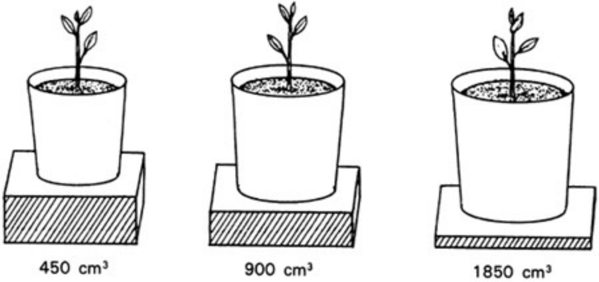
- comparing multiple populations
- quantify the differences between populations



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

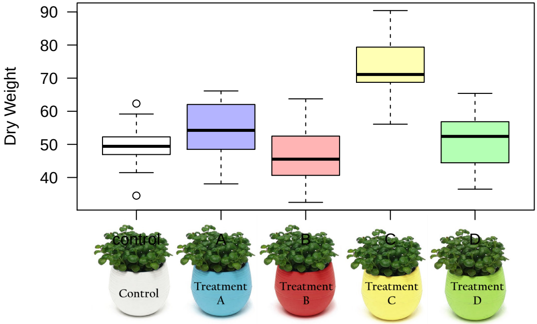
Plants productivity (dry mass in grams) related to fertilizer treatment

- 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

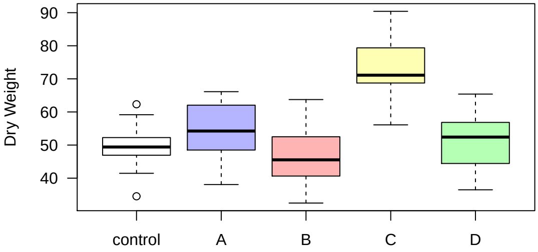
- t-test is probably not optimal:
- We would need to do all possible pairs.
- We might get spurious differences just by chance. Why?



Comparing multiple groups - ANOVA

Analysis of Variance (ANOVA):

- statistical test for testing for differences among >2 groups
- ANOVA and t-test are identical when there are 2 groups
- one factor/group/category (*One-way ANOVA*)



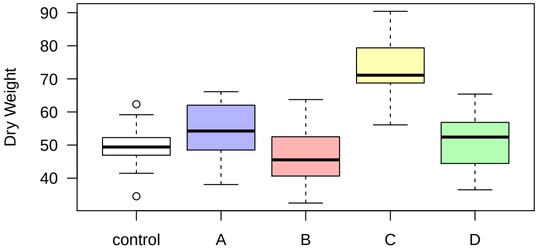
Comparing multiple groups - ANOVA

Analysis of Variance (ANOVA):

- statistical test for testing for differences among >2 groups
- ANOVA and t-test are identical when there are 2 groups
- one factor/group/category (*One-way ANOVA*)

Assumption:

- data are normally distributed



Comparing multiple groups - ANOVA

Analysis of Variance (ANOVA):

- statistical test for testing for differences among >2 groups
- ANOVA and t-test are identical when there are 2 groups
- one factor/group/category (*One-way ANOVA*)

Assumption:

- data are normally distributed

Hypotheses:

?

Comparing multiple groups - ANOVA

Analysis of Variance (ANOVA):

- statistical test for testing for differences among >2 groups
- ANOVA and t-test are identical when there are 2 groups
- one factor/group/category (*One-way ANOVA*)

Assumption:

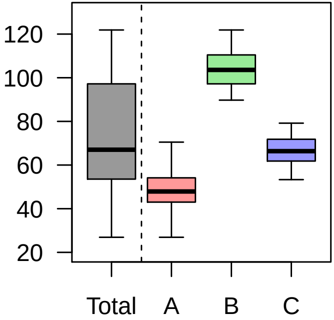
- data are normally distributed

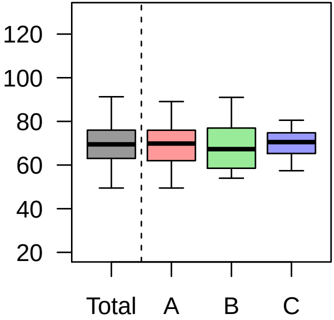
Hypotheses:

- 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

ANOVA explained

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

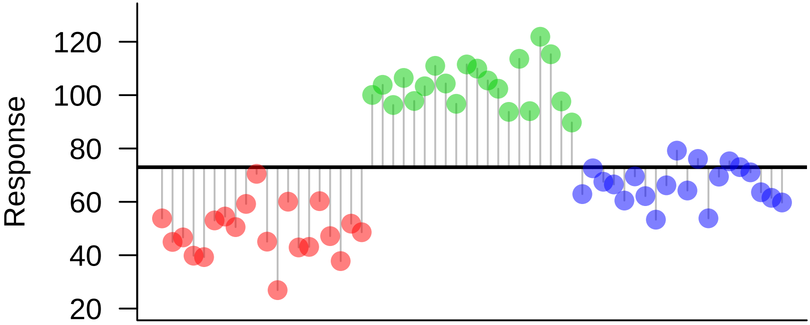




ANOVA and the Sums of Squares

- *Total* sums of squares (SS_T)

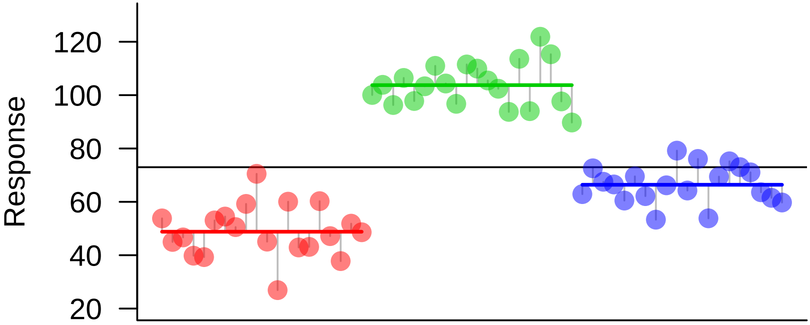
$$SS_T = \sum (x - \bar{x})^2$$



ANOVA and the Sums of Squares

- *Within-sample* sums of squares (SS_W)
- add up the within sample SS

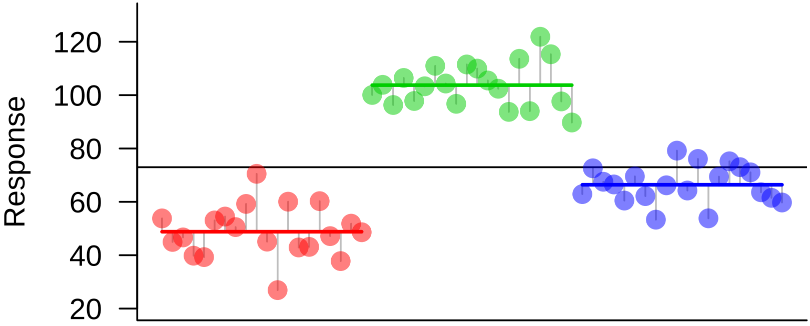
$$SS_W = \sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2 + \sum(x_3 - \bar{x}_3)^2$$



ANOVA and the Sums of Squares

- *Within-sample* sums of squares (SS_W)
- more generally (g is the number of groups)

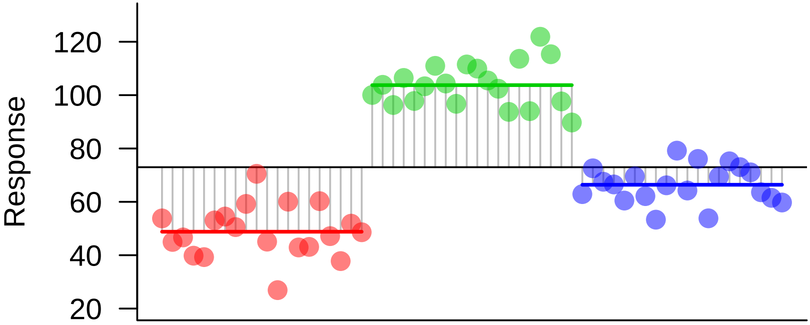
$$SS_W = \sum_g \sum_i (x_{ig} - \bar{x}_g)^2$$



ANOVA and the Sums of Squares

- *Between-sample* sums of squares (SS_B)
- add up the differences in the means

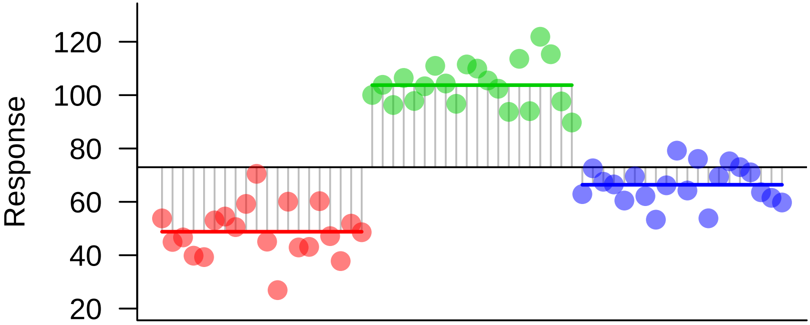
$$SS_B = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2$$



ANOVA and the Sums of Squares

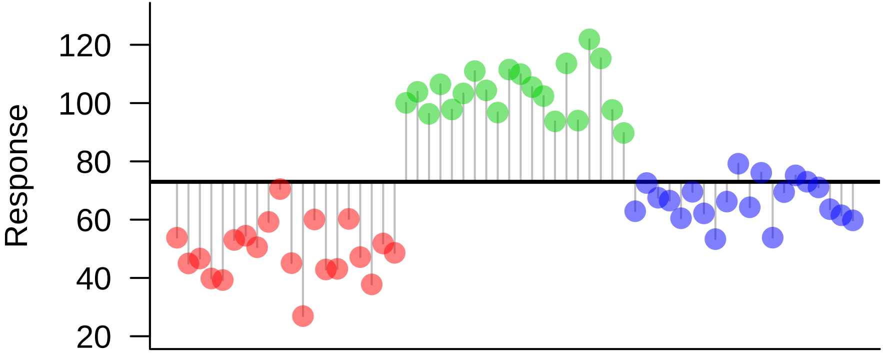
- *Between-sample* sum of squares (SS_B)
- more generally (g is the number of groups)

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



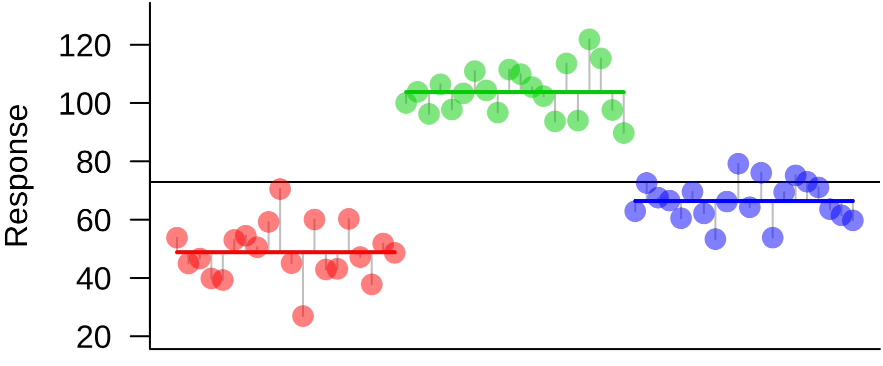
ANOVA and the Sums of Squares

Total:



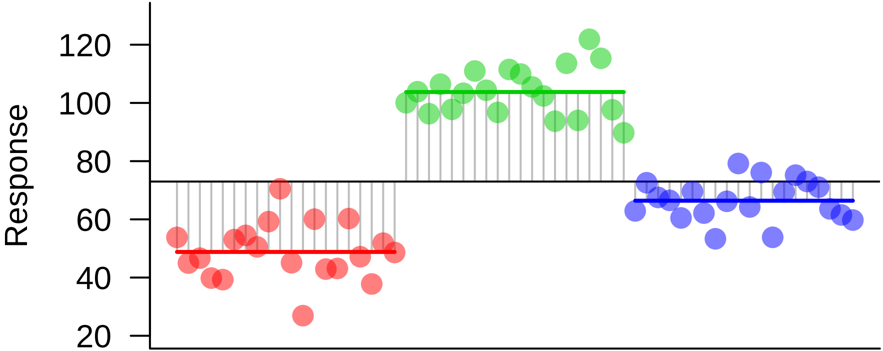
ANOVA and the Sums of Squares

Within group:



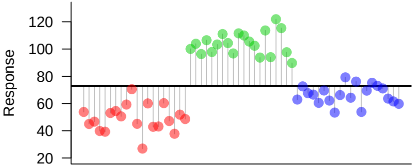
ANOVA and the Sums of Squares

Between group:

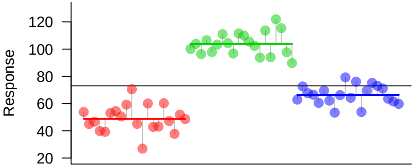


ANOVA and the Sums of Squares

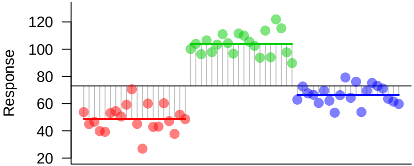
$$SS_T = \sum (x - \bar{x})^2$$



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



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



ANOVA degrees of freedom

If we define the following:

- n is the total sample size (number of observations)
- g is the number of groups/samples

ANOVA degrees of freedom

If we define the following:

- n is the total sample size (number of observations)
- g is the number of groups/samples

Then the degrees of freedom (df) are:

- Total: $df_T = n - 1$
- Within: $df_W = g - 1$
- Between: $df_B = n - g$

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: $MS_B = SS_B/df_B$

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

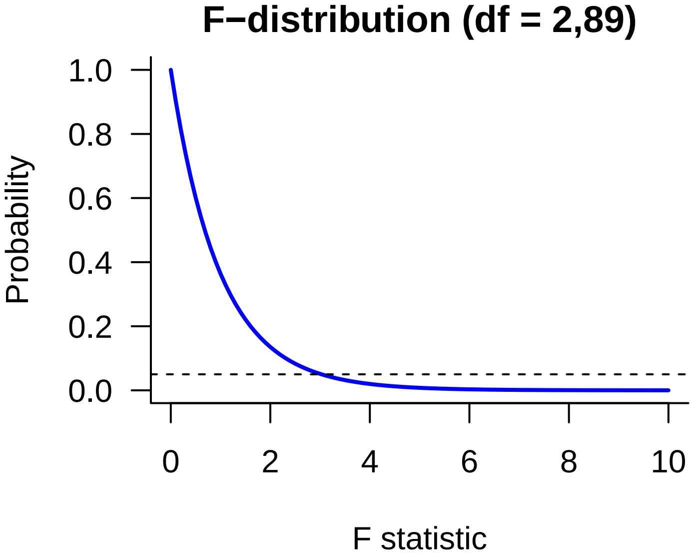
$$F = \frac{MS_B}{MS_W}$$

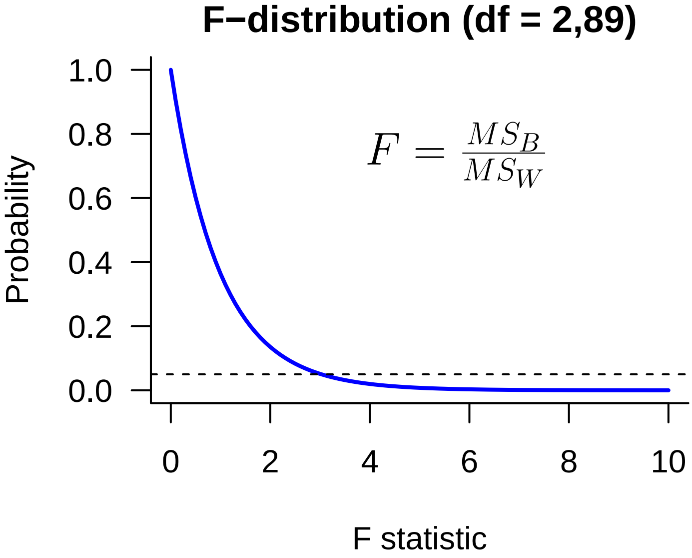
ANOVA the statistical test

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	–		

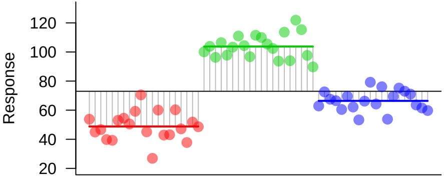
- p is the probability of observing the F statistic with a given degrees of freedom if the null hypothesis is true:
  - null hypothesis is 'no difference between the means'
  - based on the F-distribution


F−distribution (df = 2,89)
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
F statistic

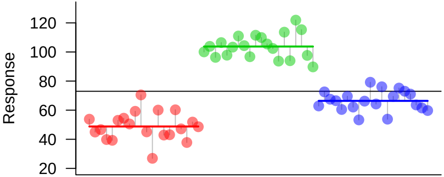


F−distribution (df = 2,89)
1.0
0.8
0.6
0.4
0.2
0.0
Probability
$F = \frac{MS_B}{MS_W}$
0
2
4
6
8
10
F statistic

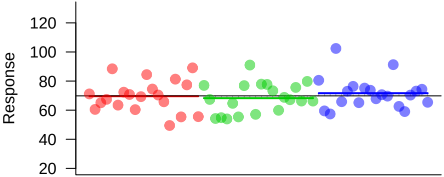
ANOVA and the Sums of Squares



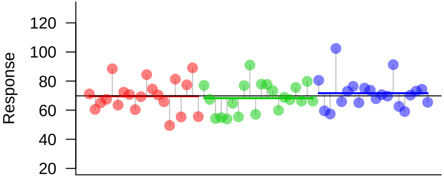
$$F = \frac{MS_B}{MS_W}$$



ANOVA and the Sums of Squares



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



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:

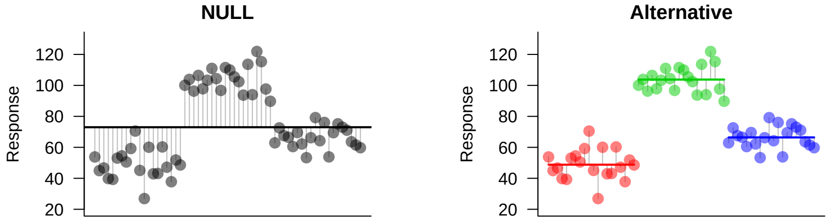
- Instead we conduct *Post-hoc* testing
  - Tukey Honest Significant Difference test (Tukey HSD)
  - accounts for multiple tests being conducted
  - calculation of a t-statistic
  - a pair, so degrees of freedom is 1
  - 5% critical value for $df = 1$ is 4.303
    - if $t > 4.303$ then $p < 0.05$

	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

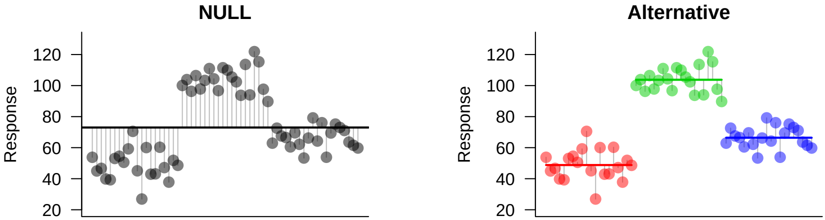
- null hypothesis:
  - no difference between the samples
  - data are from the same population
- alternative hypothesis:
  - sample means are different
  - data from the different populations



ANOVA Recap

Comparing differences between >2 groups using ANOVA

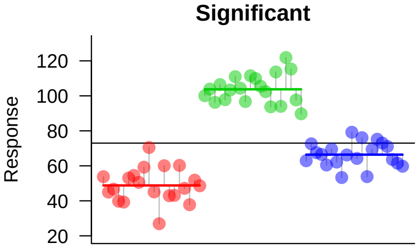
Source of variation	SS	df	MS	F	p
Between	SS_B	df_B	MS_B	$\frac{MS_B}{MS_W}$	
Within	SS_W	df_W	MS_W		
Total	SS_T	df_T	–		

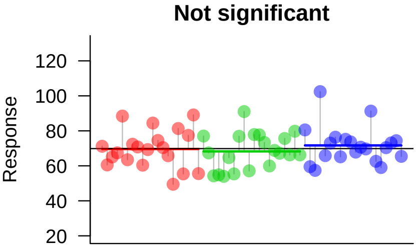


ANOVA Recap

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