Week 8: Differences among more than two samples

Session 2

Spring 2020

iClicker Question 1

I have data on salamander snout to vent length for male, female, and sex-undetermined individuals. What kind of plot could I use to assess normailty of the data:

A `scatterplot`

B `barplot`

C `histogram`

D `density plot`

E `dendrogram`

iClicker Question 2

I have data on salamander snout to vent length for male, female, and sex-undetermined individuals. My null hypothesis is that there is no difference in SVL among the sexes.

What kind of plot could I use to visually assess my null hypothesis:

A `scatterplot`

B `barplot of all data`

C `scatterplots of each sex`

D `barplots of each sex`

E `boxplots of each sex`

iClicker Question 3

I have data on salamander snout to vent length for male, female, and sex-undetermined individuals. The data were collected at two sites: A and B. I think salamanders should be bigger at site A than site B.

- I know about the follwing methods:
  - paired t-test
  - two-sample t-test with directional hypothesis
  - two-sample t-test with nondirectional hypothesis
  - one-way ANOVA
  - two-way ANOVA

How many of the above tests would help answer my question about sites A and B?

A 1

B 2

C 3

D 4

E 5

Announcements

Course moving to ‘online’ format, most likely implemented via Zoom.

Please take a moment to familiarize yourself with Zoom.

It is fairly simple and intuitive to use, but it would be advantageous to have it set up prior to our first remote session on Mar 24th.

Today

- One-way ANOVA recap
- ANOVA intuition
- Two-way ANOVA
- Begin statistical analysis of salamanders

One-way ANOVA recap

I'll ask you a question: What do you know about one-way ANOVA?

- When is it useful?
- What is the null hypothesis?
- What assumptions does it include?

ANOVA bird's eye view

- **Global** null and alternative hypothesis
- **Global** significance test
  - What does it tell us?
  - What doesn't it tell us?

One primary objective of an ANOVA is to quantify evidence that breaking up our observations into groups *improves* our description, relative to the null model that all data come from the same group.

Null hypothesis is represented by the Total Sum of Squares: SS_T.

The alternative hypothesis is represented by the Within- and Between-group Sums of Squares: SS_W and SS_B.

What is our criterion for how model improvement?

ANOVA intuition

- Sum of square terms are a way to quantify **variability**.
- Remember that SS_T is equal to the sum of SS_B and SS_W.
  - The within- and between-group sumes of square terms are calculated from different numbers of observations, so they are not directly comparable.
- We **normalize** the sums of squares by their **degrees of freedom**, which produces **mean squares** terms: MS_W and MS_B
  - This allows us to directly compare variability within- and between-groups.
- What would we expect to observe if variability was similar within- and between-groups?

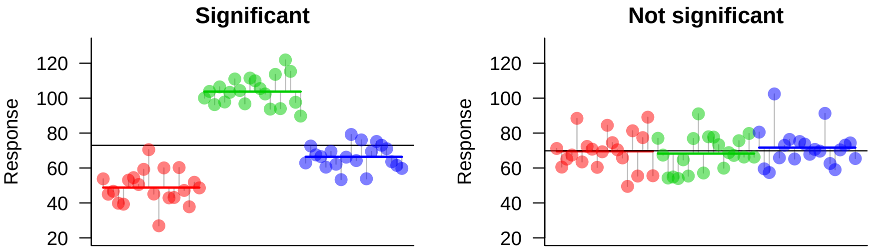
Anova intuition

- If our data were totally random, i.e. there was no difference between groups, we would expect MS_B and MS_W to be approximately equal.
  - Their ratio $\frac{SS_B}{SS_W}$ would be approximately 1.
  - Breaking our data into groups would not **improve** our model.
- If partitioning the data into groups **improved** our description, we would expect the within-group variability to be less than the variability between groups.
- Their ratio $\frac{SS_B}{SS_W}$ would be greater than 1.
- The F-statistic is the mean squares ratio. Higher values indicate greater model improvement from breaking data into groups.

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

Two-way ANOVA

Let's use a grazing example:

- Grazing in two sites: upper and lower
- Grazing with three grass heights: low, mid, and high

Two-way ANOVA

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

```r
head(graze)
```

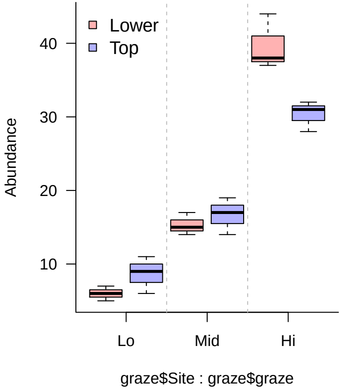
```
##   graze Site Abundance
## 1    Lo  Top         9
## 2    Lo  Top        11
## 3    Lo  Top         6
## 4   Mid  Top        14
## 5   Mid  Top        17
## 6   Mid  Top        19
```

Two-way ANOVA

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

```
head(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
```

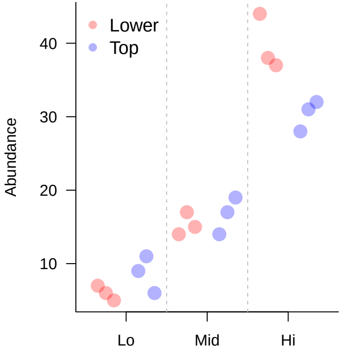


Two-way ANOVA

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

```
head(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
```



Group Exercise - *salamANOVA*

We will conduct three analyses using the *salamANOVA*. We are interested in whether salamander snout-to-vent length (SVL) varies by sex and/or site. The data look like this:

```r
sals = read.csv("mander_anova.csv")
head(sals)
```

```
##   Collector Year Season Site SVL Total_length    Sex
## 1     Chris 2014   Fall    A  36           72 female
## 2     Chris 2014   Fall    A  46           83 female
## 3     Chris 2014   Fall    A  42           89 female
## 4     Chris 2014   Fall    A  34           75 female
## 5     Chris 2014   Fall    A  37           80 female
## 6     Chris 2014   Fall    A  40           79 female
```

Group Exercise - *salamANOVA*

I can use the following syntax to conduct a 1-way anova of total length explained by collector in R:

```
fit1 = lm(sals$SVL ~ sals$Collector)

anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: sals$SVL
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## sals$Collector   2 4788.5 2394.24  110.58 < 2.2e-16 ***
## Residuals      270 5846.2   21.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Group Exercise - *salamANOVA*

I can use the following syntax to conduct a 2-way anova of total length explained by collector and site in R:

```
fit2 = lm(sals$SVL ~ sals$Collector + sals$Site)

anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: sals$SVL
##                 Df Sum Sq Mean Sq  F value  Pr(>F)
## sals$Collector   2 4788.5 2394.24 112.2247 < 2e-16 ***
## sals$Site        3  149.9   49.98   2.3426 0.07351 .
## Residuals      267 5696.3   21.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Group Exercise - *salamANOVA*

NOTE: You may not need to use all of the data in these analyses.
Remember our data-recording concepts from earlier in the course. We recorded information in our data that might be relevant at some point, but that might not be directly needed in every analysis we consider.

- `Site`: there are four sites (`P1A`, `P1B`, `P2A`, `P2B`)
- `Sex`: `M` (male) and `F` (female)
- `SVL`: the snout-to-vent length in mm

Exercise - *salamANOVA*

Analysis 1: Does SVL vary by sex?

- What is the null hypothesis?
- Make a plot to visualize the hypothesis.
- What statistical test will you use to test H_0?
- What is the:
  - test statistic for this particulat test (e.g., t, F, etc)
  - degrees of freedom (calculate this)
  - significance level
- Conduct the analysis:
  - what is the value of the test statistic
  - what the p-value
- Write a short paragraph reporting the conclusion, use values from the statistical test to suppo, supported by the results from the test.

Exercise - *salamANOVA*

Analysis 2: Does SVL vary by site?

- What is the null hypothesis?
- Make a plot to visualize the hypothesis.
- What statistical test will you use to test H_0?
- What is the:
  - test statistic for this particulat test (e.g., t, F, etc)
  - degrees of freedom (calculate this)
  - significance level
- Conduct the analysis:
  - what is the value of the test statistic
  - what the p-value
- Write a short paragraph reporting the conclusion, use values from the statistical test to suppo, supported by the results from the test.

Exercise - *salamANOVA*

Analysis 3: Does SVL vary by sex and/or site?

- What is the null hypothesis?
- Make a plot to visualize the hypothesis.
- What statistical test will you use to test H_0?
- What is the:
  - test statistic for this particulat test (e.g., t, F, etc)
  - degrees of freedom (calculate this)
  - significance level
- Conduct the analysis:
  - what is the value of the test statistic
  - what the p-value
- Write a short paragraph reporting the conclusion, use values from the statistical test to suppo, supported by the results from the test.

Exercise - *salamANOVA*

Assignment: Statistical analysis of variation in salamnder SVL.

- Write a report with four sections:
  1. Analysis 1
  2. Analysis 2
  3. Analysis 3
  4. Reflection: how does analysis 3 compare to analyses 1 and 2?
- Sections 1 to 3 sould report on each of the prompts in the previous slides.
- Section 4 is an opportunity to demonstrate your undertanding of the material covered over the previous weeks.