Statistical Models in R: Day 1

Analysis of Variance (ANOVA) in R

Tiffany Timbers
Applied Statistics and Data Science Group
UBC Statistics

February 6, 2017

ANOVA - when to use it?

Types of data

- response/dependent variable (Y) is quantitative
- examples of quantitative variables:
  - height
  - salary
  - number of offspring
- your explanatory/independent variable(s) (X's) are categorical
- examples of categorical variables:
  - eye color
  - sex
  - genotype at a given locus

Examples of Cases where ANOVA would be used

1. Does diet has an effect on weight gain?
   - response variable = weight gain (*e.g.*, kg)
   - explanatory variable = type of diet (*e.g.*, low vs. medium vs. high sugar)
2. Does the type sexual relationship practiced influence the fitness of male Red-winged Blackbirds?
   - response variable = fitness of male bird (*e.g.*, # eggs laid)
   - explanatory variable = sexual relationship (*e.g.*, monagamy vs. polygamy)

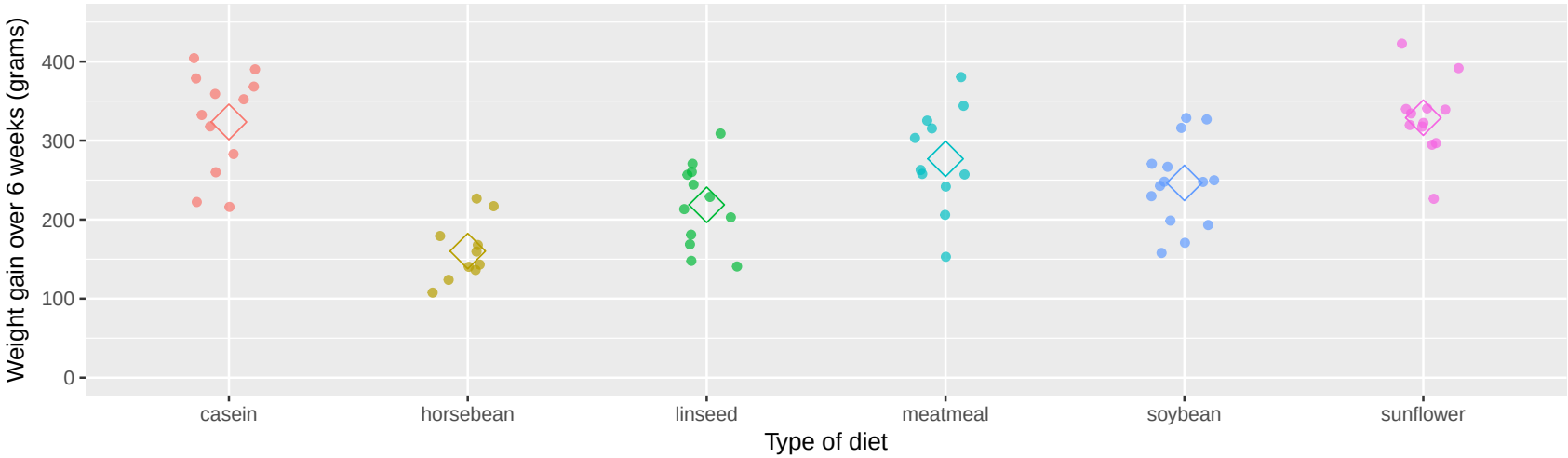
ANOVA - how it works?

Example case:

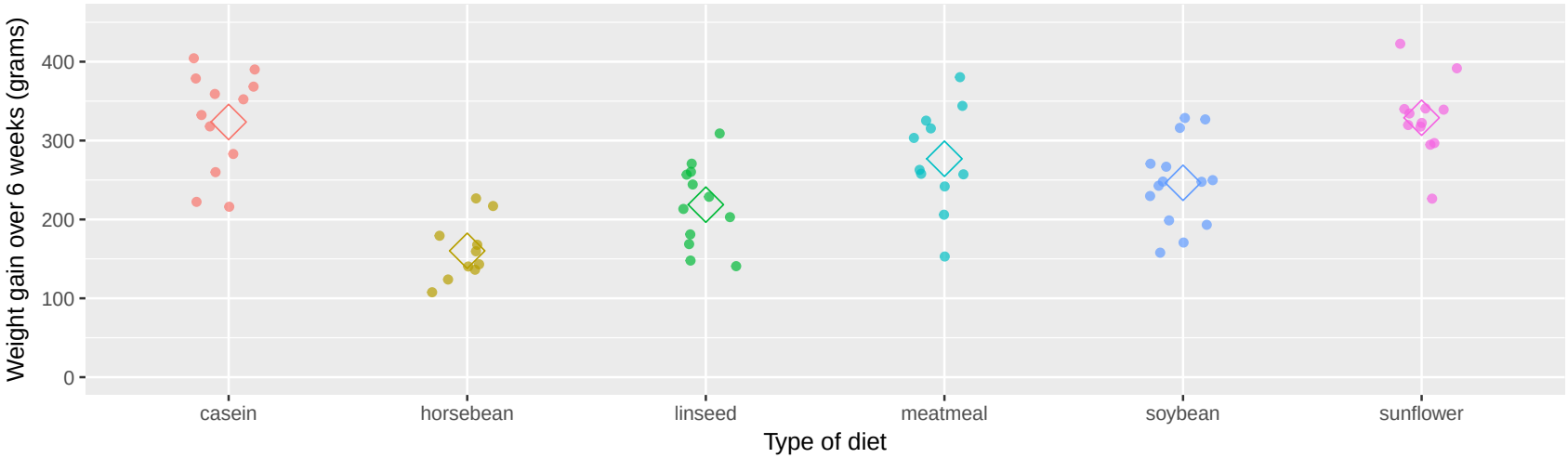
- Does diet has an effect on chick weight after 6 weeks?
- response variable == chick weight (grams)
- explanatory variable == type of diet (4 levels)

$H_0 : \mu_{caesin} = \mu_{horsebean} = \mu_{linseed} = \mu_{meatmeal} = \mu_{soybean} = \mu_{sunflower}$

H_A : at least one group's population mean differs from that of the other groups



The gist of the math behind it



To calculate a test statistic (F-statistic), we compare the **between** group variation with the **within** group variation

$$F = MSB/MSW$$

where, MSB = Mean Square Between & MSW = Mean Square Within
Essentially, if there is greater variation between the groups than within the groups we will get a large test statistic value (and correspondingly, a small p-value) and reject that null hypothesis (H_0: population means of all groups are equal).

Want to know more?

Watch this excellent series of videos from Khan Academy where they perform ANOVA by hand for a simple case:
https://www.khanacademy.org/math/statistics-probability/analysis-of-variance-anova-library

Key assumptions / rules of thumb

- ANOVA is robust to the non-normality of sample data
- Balanced ANOVA (equal sample size between groups) is robust to uneuqal variance
- ANOVA is sensitive to sample indpendence

Syntax for ANOVA in R

Laying out your dataset

- Most statistical functions in R work best with a "tidy" dataset
  - tidy data def'n: *"each variable is a column, each observation is a row"*
- For our case (chick feed example), this means 2 columns, one for the response variable and one for the explanatory variable:

weight	feed
143	horsebean
295	sunflower
320	sunflower
222	casein
344	meatmeal
216	casein

for more information on tidy data see:
http://vita.had.co.nz/papers/tidy-data.pdf

The aov function

- To perform ANOVA in R we can use the `aov` function
- `aov` requires the following arguments:
  - `formula`
  - `data`

Here is an example call to `aov` using our chick feed case example:

```
chick_feed_model <- aov(weight ~ feed, data = chickwts)
```

To get the results from the ANOVA analysis performed by `aov` in a nice tidy data frame, I recomend using the `tidy` function from R's `broom` package:

```
library(broom)
broom::tidy(chick_feed_model)
```

```
##         term df    sumsq    meansq statistic     p.value
## 1       feed  5 231129.2 46225.832   15.3648 5.93642e-10
## 2 Residuals 65 195556.0  3008.554        NA          NA
```

summary - another way to get results from aov

Traditionally the summary function was used. This provides a nice print output, but unlike tidy, the results are not contained in a data frame, and thus more tricky to access:

```
summary(chick_feed_model)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## feed          5 231129   46226   15.37 5.94e-10 ***
## Residuals    65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```