Lesson 23. The One-Way ANOVA Model – Part 1

Note. In Part 2 of this lesson, you can run the R code that generates the plots and outputs in here Part 1.

1 Overview

- Suppose we have:
  1. One quantitative response variable
  2. One categorical explanatory variable that breaks the sample into groups
     - We refer to the groups as **treatments** or **levels**
- Key questions:
  1. How strong is the evidence that the treatment makes a difference in the response?
  2. If there is a difference due to treatment, how big is it?

Example 1. A study was designed to compare the effect of three different high-protein diets on weight gain in baby rats. The data is stored in `FatRats` in our textbook data library `Stat2Data`.

The subjects for the study were 30 baby rats. Each was fed a high-protein diet from one of three sources: beef, cereal, or pork. Their weight gains were recorded in grams. We would like to test whether average weight gain differs from protein source.

a. Is this an observational study or an experiment?

b. What are the "treatments"?

c. In R, we load and preview the data:

```
library(Stat2Data)
data(FatRats)
head(FatRats)
```

Here is the output:

A data.frame: 6 × 3

	Gain	Protein	Source
	<int>	<fct>	<fct>
1	73	Hi	Beef
2	102	Hi	Beef
3	118	Hi	Beef
4	104	Hi	Beef
5	81	Hi	Beef
6	107	Hi	Beef

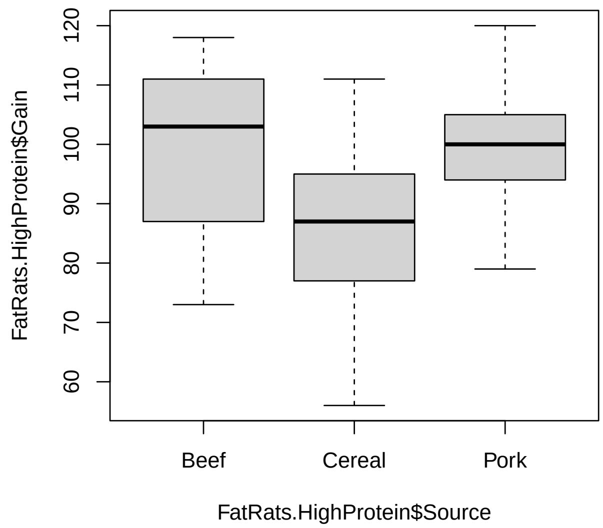
Next, we create a new dataframe, keeping only the rats who got a high-protein diet:

```
FatRats.HighProtein <- FatRats[FatRats$Protein == 'Hi', ]
```

We make boxplots to visualize the weight gains grouped by protein source:

```
boxplot(FatRats.HighProtein$Gain ~ FatRats.HighProtein$Source)
```

Here is the output:



We also use R to compute the mean weight gain for each group:

```
ybar.k <- tapply(FatRats.HighProtein$Gain, FatRats.HighProtein$Source, mean)
ybar.k
```

Here is the output:

Beef: 100 **Cereal:** 85.9 **Pork:** 99.5

d. What are the key questions we are trying to answer?

2 The one-way ANOVA model

- We need:
  - One quantiative response variable
  - One categorical explanatory variable with K values
- The model:

$$Y \quad = \quad \mu \quad + \quad \alpha_k \quad + \quad \varepsilon \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

- Parameter estimates:

$$\hat{\mu} = \bar{y} \qquad\qquad \hat{\alpha}_k = \bar{y}_k - \bar{y}$$

- The ANOVA table:

Source	DF	Sum of Squares	Mean Square	F-Statistic
Groups				
Error				
Total				

2.1 A brief aside: haven't we seen ANOVA before?

- We have seen ANOVA before, in the context of linear regression
  - We used an ANOVA table to determine whether an overall regression model was effective or not
- In this (and subsequent) lessons, ANOVA refers not only to the table itself but also the model that we will use
  - Note that the one-way ANOVA model requires one categorical explanatory variable

Example 2. Continuing with the `FatRats` setting from Example 1...

a. In R, we can get the parameter estimates as follows:

```
ybar <- mean(FatRats.HighProtein$Gain)
alpha.k <- ybar.k - ybar

ybar
alpha.k
```

Here is the output:

95.1333333333333
Beef: 4.86666666666666 **Cereal:** -9.23333333333333 **Pork:** 4.36666666666666

b. We can get the ANOVA table as follows:

```
test <- aov(FatRats.HighProtein$Gain ~ FatRats.HighProtein$Source)
summary(test)
```

Here is the output:

```
                           Df Sum Sq Mean Sq F value Pr(>F)
FatRats.HighProtein$Source  2   1280   640.0   3.346 0.0503 .
Residuals                  27   5165   191.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3 Hypothesis testing: the one-way ANOVA F-test for K groups

- Question: **Is there a difference between the means of the different groups?**
- Formal steps:

1. State the hypotheses:

2. Calculate the test statistic:

3. Calculate the p-value:
   - If the conditions for multiple linear regression hold, then the sampling distribution of the test statistic under the null hypothesis is the F-distribution with

     degrees of freedom

4. State your conclusion, based on the given significance level α:

   If we reject H_0 (p-value $\leq \alpha$):

   We see significant evidence that the mean response differs by the treatments.

   If we fail to reject H_0 (p-value $> \alpha$):

   We do not see sufficient evidence that the mean response differs by the treatments.

Example 3. Continuing with the `FatRats` setting from Examples 1 and 2...

Do we see significant statistical evidence that the mean weight gain differs by protein source? Using the output above, perform a one-way ANOVA F-test.

4 Coming next...

- Conditions under which an ANOVA model is appropriate
- If there is a difference due to treatment, how big is it?