Lesson 5. Hypothesis Testing, cont. – Part 1

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

1 Overview

- Last lesson: t-test for one population mean
  - Question: Is an unknown population mean different from a specific value?
- This lesson: another hypothesis test for a different question...
  - Question: Is there a difference between the means from two or more populations or **groups**?
- **Analysis of variance (ANOVA)** is an approach to comparing the means of different groups by examining variation in the data
- First, a motivating example

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 course 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.

⚠ We will often use datasets from our course textbook, STAT2. Conveniently, there is a R package that contains all these datasets. You can install it by running the code below in an empty code cell. You only need to do this once!

```
install.packages('Stat2Data')
```

- 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

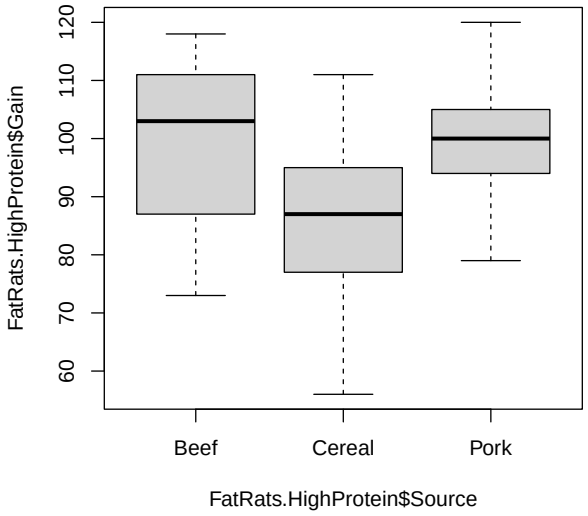
- Next, we create a new dataframe, keeping only the rats who got a high-protein diet:

```
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 compute the mean weight gain for each group:

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

Here is the output:

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

- What are the key questions we are trying to answer?

2 The one-way ANOVA table

- Let n be the number of observations, and k be the number of comparison groups

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

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

- In R, we can get the one-way 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
```

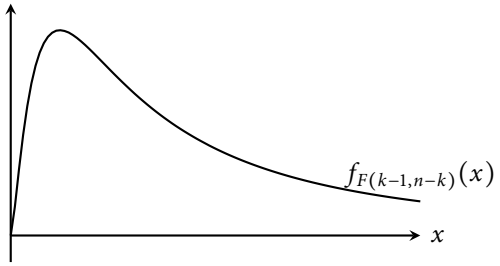
3 One-way ANOVA F-test for k groups

- Question: **Is there a difference between the means μ_i of the different groups $i = 1, \ldots, k$?**
- Formal steps:
  1. State the hypotheses:
  2. Calculate the test statistic:

3. Calculate the p-value:

   If H_0 is true, the test statistic F follows an F-distribution with $k-1$ and $n-k$ degrees of freedom.

   p-value =



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

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

   At the significance level of α, we reject the null hypothesis. We see evidence that the mean differs by the groups.

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

   At the significance level of α, we fail to reject the null hypothesis. We do not see evidence that the mean differs by the groups.

   The underlined parts above should be rephrased to correspond to the context of the problem.

- Technical conditions to check – this test may be used if the following conditions hold:

  1. Each population (group) has the same standard deviations
     - Check:
  2. Each population (group) is normally distributed
     - Check:
  3. After accounting for group membership, responses are independent
     - Check:

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.

Check whether each of the three conditions for the one-way ANOVA F-test appears to be met.

- We can compute the standard deviations for each group using `tapply`:

  ```
  tapply(FatRats.HighProtein$Gain, FatRats.HighProtein$Source, sd)
  ```

  Here is the output:

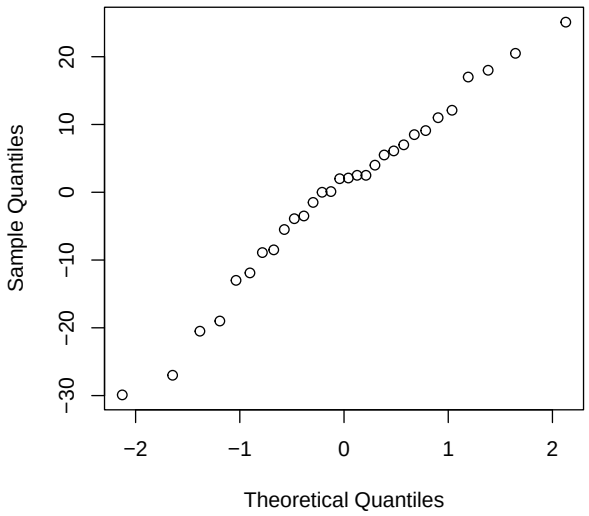
  Beef: 15.136416719657 **Cereal:** 15.0218359582161 **Pork:** 10.9163485958752

  Is the standard deviation condition met?

- We can create a normal Q-Q plot of the residuals values as follows:

```
qqnorm(residuals(test))
```

  Here is the output:



  Is the normality condition met?

- After accounting for group membership, are the responses independent?