Chapters covered: Chapter 14

- Problem 1: Textbook Problem 14.6 (ANOVA and box plots) page 679
- Problem 2 : Textbook Problem 14.36 (Good friends and marital status) page 704. For c. Use the following R command to verify the P-value.

  ```
  pf(____, df1= _____, df2 = _______, lower.tail=FALSE)
  ```

  Additional problem: d. How large an F test statistic is needed to get a P-value $= 0.05$ in this ANOVA? Include your R command and output.
- Problem 3: Textbook Problem 14.62 (Multiple choice: ANOVA variability)
- Problem 4 : The following R commands and outputs show scores on the first quiz (maximum score of 10 points) for eighth-grade students in an introductory level French course. The instructor grouped the student in the course as follows:
  - Group 1: Never studied foreign language before but have good English skills
  - Group 2: Never studied foreign language before and have poor English skills
  - Group 3: Studied at least one other foreign language

Quiz scores			
	Group 1	Group 2	Group 3
	4	1	9
	6	5	10
	8		5
mean	$\bar{y}_1 = 6$	$\bar{y}_2 = 3$	$\bar{y}_3 = 8$
standard deviation	$s_1 = 2$	$s_2 = 2.8$	$s_3 = 2.6$

```
> test <-aov(scores~group, data=quiz)
> summary(test)
            Df Sum Sq Mean Sq F value Pr(>F)
group        2     30      15     2.5  0.177
Residuals    5     30       6
```

Use R commands and outputs above to answer part a. through h. Assume that quiz scores are normally distributed.

a. State the assumptions of ANOVA.

b. Introduce notations, state the null and alternative hypotheses.

c. Identify the value of test-statistic from R output above.

d. Identify P-value and draw conclusion at $\alpha = 0.05$.

e. Is the result statistically significant? Why or why not?

f. Is the result practically significant? Why or why not? (answer may vary).

g. Do you think any follow-up procedure such as Tukey's HSD is necessary? Why or why not?

h. Suppose that the first observation in the second group was actually 9 not 1. Then the standard deviations are the same as reported in the table, but the sample means are 6, 7, and 8 rather than 6, 3, 8. Do you think thee F-test statistic would be larger, the same, or smaller? Explain the reason without doing any calculation.

- R Problem : The 2014 General Social Survey asked randomly selected 1475 subjects how many hours per day they watched TV, on average. We want to test whether there are differences in population means according to the marital status of the subject. The data set `TVdata` contains three variables (TVhours, status). Use the following R command to import the data.

```
dat<-read.csv("http://users.stat.umn.edu/~parky/TVdata.csv")
```

a. Identify the explanatory variable and response variable.

b. Construct a side-by-side boxplot. Describe the distributions.

c. Define the parameters of interest and state the null and alternative hypotheses.

d. Use R to construct ANOVA table. Copy and paste your R output. Identify p-value and draw a conclusion at $\alpha = 0.05$.

e. The following R output shows the Tukey's multiple comparison results with overall confidence level of 0.95. Identify pair(s) with significant different mean TV hours and interpret the interval.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = TVhours ~ status, data = TVdata)

$`status`
                        diff         lwr       upr     p adj
married-divorced -0.2096767 -0.60942405 0.1900706 0.4352694
single-divorced   0.4131636 -0.01853188 0.8448591 0.0641301
single-married    0.6228404  0.28193935 0.9637414 0.0000574
```