HW4: Sections 5.3, 5.4

Your Name Here

The code below just loads some packages and makes it so that enough digits are printed that you won't get confused by rounding errors.

```
library(dplyr) # functions like summarize
library(ggplot2) # for making plots
library(mosaic) # convenient interface to t.test function
```

```
## Warning: package 'mosaic' was built under R version 3.5.2
```

```
## Warning: package 'ggformula' was built under R version 3.5.2
```

```
library(readr)
library(gmodels)

options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Problem 1: Adapted from Sleuth3 5.18

A randomized experiment was conducted to estimate the effect of a certain fatty acid (CPFA) on the level of a certain protein in rat livers. Only one level of the CPFA could be investigated in a day's work, so a control group (no CPFA) was investigated each day as well. The following R code reads in the data.

```
cpfa <- read.csv("http://www.evanlray.com/data/sleuth3/ex0518_fatty_acid.csv")
head(cpfa)
```

```
##   Protein Treatment  Day TrtDayGroup
## 1     154    CPFA50 Day1      Group1
## 2     177    CPFA50 Day1      Group1
## 3     174    CPFA50 Day1      Group1
## 4     164   CPFA150 Day2      Group2
## 5     192   CPFA150 Day2      Group2
## 6     159   CPFA150 Day2      Group2
```

```
cpfa %>% distinct(Treatment, Day, TrtDayGroup)
```

```
##    Treatment  Day TrtDayGroup
## 1     CPFA50 Day1      Group1
## 2    CPFA150 Day2      Group2
## 3    CPFA300 Day3      Group3
## 4    CPFA450 Day4      Group4
## 5    CPFA600 Day5      Group5
## 6    Control Day1      Group6
## 7    Control Day2      Group7
## 8    Control Day3      Group8
## 9    Control Day4      Group9
## 10   Control Day5     Group10
```

Display 5.21 in the book shows the organization of how the data were collected, but basically there are 6 treatments (recorded in the `Treatment` variable in the data set, with values CPFA50, CPFA150, CPFA300, CPFA450, CPFA600, and Control), and the experiment was run over the course of 5 days. In the data frame, the `TrtDayGroup` records a unique combination of values for the Treatment and the Day. For example, Group1 is for the three observations that were made for the CPFA50 treatment on Day1, and Group2 is for the three observations for the CPFA150 treatment on Day2, and so on. There are 10 groups total, 5 for the 5 CPFA treatments and 5 more for the control treatment which was run on all 5 days. The assignment of each combination of a treatment and a day to one of the 10 combinations is shown in the R code output above.

(a) Fit a model that uses the `TrtDayGroup` as the explanatory variable to estimate a separate mean protein level for each of the 10 treatment-day combinations. Conduct the ANOVA F test to see whether these 10 groups have equal means. State your hypotheses clearly using symbols for the 10 means and a written sentence explaining the interpretation of each hypothesis in context; interpret the results of the test in context as well.

(b) Fit a reduced model that uses the `Treatment` as the explanatory variable to fit a separate mean for each of the 6 treatments. Conduct the ANOVA F test to compare the full model in part a to the reduced model with 6 means. State the hypotheses the hypotheses that are being tested clearly using symbols for the 10 means in the full model and a written sentence explaining the interpretation of each hypothesis in context; interpret the results of the test in context as well.

Problem 2: Sleuth3 5.25

The R code below reads in data with annual incomes as of 2005 for a random sample of 2584 Americans who were selected for the National Longitudinal Survey of Youth in 1979 and who had paying jobs in 2005. The data set also includes a code for the number of years of education that each individual had completed by 2006: <12, 12, 13-15, 16, or >16.

I have also added a new variable to the data frame called `sqrt_Income2005`, with the square root of each individual's income in 2005. The reason for this is that the ANOVA model asserts that the response variable follows a normal distribution within each group, but the incomes are skewed right. The transformed incomes come closer to following a normal distribution. We will talk more about data transformations next; for this assignment, just work with the square root of the income variable.

```r
income <- read.csv("http://www.evanlray.com/data/sleuth3/ex0525_education_income.csv")
income <- income %>%
  mutate(
    Educ = factor(Educ, levels = c("<12", "12", "13-15", "16", ">16")),
    sqrt_Income2005 = sqrt(Income2005)
  )
```

(a) Make a suitable plot of the data, showing the distribution of values of `sqrt_Income2005` separately for each level of `Educ`.

(b) Do the data provide evidence that at least one of the five groups has a different mean (of the square root of) income than the other groups? Conduct a relevant hypothesis test, clearly stating your hypotheses in terms of equations involving some of the group means and written sentences explaining what each hypothesis means in context. Also interpret your results in context.

(c) Do the data provide evidence that there is a difference in the mean (of the square root of) income for people with an undergraduate college degree ("16") and people with graduate level study (">16")? Conduct a relevant hypothesis test, clearly stating your hypotheses in

terms of equations involving some of the group means and written sentences explaining what each hypothesis means in context. Also interpret your results in context.

(d) Do the data provide evidence that there is a difference in the mean (of the square root of) income for people with less than an undergraduate degree (“<12”, “12”, or “13-15”)? Conduct a relevant hypothesis test, clearly stating your hypotheses in terms of equations involving some of the group means and written sentences explaining what each hypothesis means in context. Also interpret your results in context.