HW4: Section 5.3, 5.4, 6.3, and 6.4 (F tests and multiple comparisons)

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
```

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

```
library(ggplot2) # for making plots
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
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") %>%
  mutate(
    Day = as.character(Day),
    TrtDayGroup = ifelse(
      Treatment == "Control",
      paste0("Group", sprintf("%02d", as.numeric(substr(Day, 4, nchar(Day))))),
      paste0("Group", sprintf("%02d", 5 + as.numeric(substr(Day, 4, nchar(Day)))))
    )
  ) %>%
  arrange(TrtDayGroup)

head(cpfa)
```

```
##   Protein Treatment  Day TrtDayGroup
## 1     157   Control Day1     Group01
## 2     165   Control Day1     Group01
## 3     150   Control Day1     Group01
## 4     186   Control Day2     Group02
## 5     206   Control Day2     Group02
## 6     195   Control Day2     Group02
```

Display 5.21 in the book, included below, shows the organization of how the data were collected (the figure will show up in the pdf if you knit the document).

DISPLAY 5.21 Levels of protein (×10) found in rat livers

	Treatment					
Day	CPFA 50	CPFA 150	CPFA 300	CPFA 450	CPFA 600	Control
1	154, 177, 174					157, 165, 150
2		164, 192, 159				186, 206, 195
3			157, 159, 124			192, 202, 216
4				160, 152, 141		190, 187, 160
5					147, 152, 158	191, 188, 199

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 CPFA treatment group names, the number indicates the dose of CPFA given to the rat; for example, rats in the CPFA50 group received 50 units (the book doesn't tell us what the units are) of CPFA.

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 below.

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

```
##    Treatment  Day TrtDayGroup
## 1    Control Day1     Group01
## 2    Control Day2     Group02
## 3    Control Day3     Group03
## 4    Control Day4     Group04
## 5    Control Day5     Group05
## 6     CPFA50 Day1     Group06
## 7    CPFA150 Day2     Group07
## 8    CPFA300 Day3     Group08
## 9    CPFA450 Day4     Group09
## 10   CPFA600 Day5     Group10
```

(a) Fit a model that uses the `TrtDayGroup` as the explanatory variable to fit a separate mean 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.

```
mfull <- lm(Protein ~ TrtDayGroup, data = cpfa)
anova(mfull)
```

```
## Analysis of Variance Table
##
## Response: Protein
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## TrtDayGroup  9 11147.5 1238.61  7.8014 7.154e-05 ***
## Residuals   20  3175.3  158.77
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We define the following parameters:

- μ_1 is the mean protein count for the population of rats that are administered CPFA50 under conditions similar to those used in Day 1 of the experiment (corresponding to group 1)
- μ_2 is the mean protein count for the population of rats that are administered CPFA150 under conditions similar to those used in Day 2 of the experiment (corresponding to group 2)
- ...
- μ_{10} is the mean protein count for the population of rats that are not administered CPFA under conditions similar to those used in Day 5 of the experiment (corresponding to group 10)

The hypotheses are:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_{10}$: The average protein count is the same across all days and all treatment conditions.

H_A : At least one of $\mu_1, \mu_2, \ldots, \mu_{10}$ is not equal to the others. The different combinations of treatments and experiment days do not all have the same mean protein counts.

The p-value for this test is 7.15×10^{-5}. The data provide extremely strong evidence that the mean protein count is not equal across all 10 combinations of treatment group and experiment day.

(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 your hypotheses 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.

```
mred <- lm(Protein ~ Treatment, data = cpfa)
anova(mred, mfull)
```

```
## Analysis of Variance Table
##
## Model 1: Protein ~ Treatment
## Model 2: Protein ~ TrtDayGroup
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     24 7100.3
## 2     20 3175.3  4    3924.9 6.1803 0.002089 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In terms of the parameters defined in the answer to part (a), our hypotheses are:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. The mean protein count in the population of rats not given any CPFA is the same across all 5 days of the experiment.

H_A : At least one of μ_6 through μ_{10} is not equal to the others. The mean protein counts among rats not given any CPFA is different in the experimental conditions from different days.

The p-value for this test is 0.002. The data provide strong evidence that the population mean protein count was different for rats in the control groups under conditions like those on the different days the experiment was run.

(c) Print out the summary of your full model from part (a) and use the summary output to conduct a test of the claim that there is no difference in population mean protein levels between the control group and the CPFA50 group, in the "population" of rats evaluated under

conditions similar to those on day 1. Interpret your results in context in terms of strength of evidence against the null hypothesis.

```
summary(mfull)
```

```
##
## Call:
## lm(formula = Protein ~ TrtDayGroup, data = cpfa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6667  -7.5833  -0.3333   8.5000  20.3333
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          157.333      7.275  21.627 2.41e-15 ***
## TrtDayGroupGroup02    38.333     10.288   3.726 0.001334 **
## TrtDayGroupGroup03    46.000     10.288   4.471 0.000234 ***
## TrtDayGroupGroup04    21.667     10.288   2.106 0.048037 *
## TrtDayGroupGroup05    35.333     10.288   3.434 0.002625 **
## TrtDayGroupGroup06    11.000     10.288   1.069 0.297715
## TrtDayGroupGroup07    14.333     10.288   1.393 0.178852
## TrtDayGroupGroup08   -10.667     10.288  -1.037 0.312203
## TrtDayGroupGroup09    -6.333     10.288  -0.616 0.545101
## TrtDayGroupGroup10    -5.000     10.288  -0.486 0.632250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.6 on 20 degrees of freedom
## Multiple R-squared:  0.7783, Adjusted R-squared:  0.6785
## F-statistic: 7.801 on 9 and 20 DF,  p-value: 7.154e-05
```

Treatment day group 1 was for rats in the control condition on day 1, and group 6 was for rats in the CPFA50 condition on day 1. The row in the summary output labeled `TrtDayGroupGroup06` describes an estimate of the difference in population means for those conditions, as well as a hypothesis test of the claim that there is no difference in those means. Therefore, the p-value for the test we are conducting is 0.298. The data do not provide evidence of a difference in means for the "population" of rats in conditions similar to the control group on day 1 of the experiment and the "population" of rats in conditions similar to the CPFA50 group on day 1.

(d) Based on your full model, obtain a set of 5 confidence intervals for the differences in means between the control group and the corresponding treatment group that was evaluated on the same day. Please target a familywise confidence level of 95% using the Bonferroni adjustment. (You just need to write and run the code in this part.)

```
fit.contrast(mfull, "TrtDayGroup", c(1, 0, 0, 0, 0, -1, 0, 0, 0, 0), conf.int = 0.99)
```

```
##                                       Estimate Std. Error   t value
## TrtDayGroup c=( 1 0 0 0 0 -1 0 0 0 0 )      -11   10.28807 -1.069199
##                                        Pr(>|t|)  lower CI upper CI
## TrtDayGroup c=( 1 0 0 0 0 -1 0 0 0 0 ) 0.2977152 -40.27306 18.27306
## attr(,"class")
## [1] "fit_contrast"
```

```
fit.contrast(mfull, "TrtDayGroup", c(0, 1, 0, 0, 0, 0, -1, 0, 0, 0), conf.int = 0.99)
```

```
##                                         Estimate Std. Error  t value
## TrtDayGroup c=( 0 1 0 0 0 0 -1 0 0 0 )       24   10.28807 2.332798
##                                           Pr(>|t|)  lower CI upper CI
## TrtDayGroup c=( 0 1 0 0 0 0 -1 0 0 0 ) 0.03021609 -5.273062 53.27306
## attr(,"class")
## [1] "fit_contrast"
```

```
fit.contrast(mfull, "TrtDayGroup", c(0, 0, 1, 0, 0, 0, 0, -1, 0, 0), conf.int = 0.99)
```

```
##                                         Estimate Std. Error  t value
## TrtDayGroup c=( 0 0 1 0 0 0 0 -1 0 0 ) 56.66667   10.28807 5.507996
##                                             Pr(>|t|) lower CI upper CI
## TrtDayGroup c=( 0 0 1 0 0 0 0 -1 0 0 ) 2.163029e-05  27.3936 85.93973
## attr(,"class")
## [1] "fit_contrast"
```

```
fit.contrast(mfull, "TrtDayGroup", c(0, 0, 0, 1, 0, 0, 0, 0, -1, 0), conf.int = 0.99)
```

```
##                                         Estimate Std. Error  t value
## TrtDayGroup c=( 0 0 0 1 0 0 0 0 -1 0 )       28   10.28807 2.721598
##                                           Pr(>|t|)  lower CI upper CI
## TrtDayGroup c=( 0 0 0 1 0 0 0 0 -1 0 ) 0.01314121 -1.273062 57.27306
## attr(,"class")
## [1] "fit_contrast"
```

```
fit.contrast(mfull, "TrtDayGroup", c(0, 0, 0, 0, 1, 0, 0, 0, 0, -1), conf.int = 0.99)
```

```
##                                         Estimate Std. Error  t value
## TrtDayGroup c=( 0 0 0 0 1 0 0 0 0 -1 ) 40.33333   10.28807 3.920397
##                                             Pr(>|t|) lower CI upper CI
## TrtDayGroup c=( 0 0 0 0 1 0 0 0 0 -1 ) 0.0008474046 11.06027  69.6064
## attr(,"class")
## [1] "fit_contrast"
```

(e) What does it mean that your intervals from part (d) have a 95% familywise confidence level?

For 95% of samples, all 5 of the confidence intervals calculated using this procedure would contain the difference in population means they are estimating.

(f) How did R calculate the numbers labeled `Estimate` in the output for your answer to part (d)? (We discussed this on Fri, Sep 20, and related ideas on Mon, Sep 16)

A difference in population means is estimated by the corresponding difference in sample means. For example, the difference in population means for the control group and the CPFA50 treatment group under conditions similar to those on day 1 would be calculated as the difference in sample means for those two groups.

(g) How would you sum up the results of our analysis of these data? Describe the substantive findings from the experiment using your estimates and confidence intervals from part (d) to back up your conclusions. As part of your answer, address the scope of conclusions for the findings, including whether a causal association can be established and the population we can apply our findings to. Is there a minimum dose at which we have evidence that CPFA is associated with increased levels of protein? (Imagine you're writing the main paragraph for the results section in a research article about this study. For a sense of scale, it took me

6 sentences to answer this question; you will not be able to get enough detail into 1 or 2 sentences.)

This experiment gives us strong evidence of a causal relationship between CPFA dose and reduced levels of protein in rat livers, in the population of rats similar to those in this study under laboratory conditions. The treatment groups receiving doses of 300 and 600 units of CPFA had much lower mean levels of proteins in the liver than the control groups. For those groups, the estimated differences in mean protein levels from the control groups were about 57 and 40, with 95% familywise confidence intervals of [27, 86] and [11, 70] respectively. There was no evidence of a difference in mean protein levels for the control group and the treatment group receiving a dose of 50 units of CPFA. Once we account for the fact that we conducted multiple comparisons for the different treatment groups, there was also not strong evidence of a difference in mean protein levels between the control groups and the groups receiving doses of 150 or 450 units of CPFA, with 95% familywise confidence intervals of [-5, 53] and [-1, 57] for those differences respectively. However, given that there was strong evidence of increase protein levels in rats receiving doses of 300 or 600 units of CPFA, it seems likely that a dose of 450 units of CPFA is also associated with higher mean protein levels.

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.

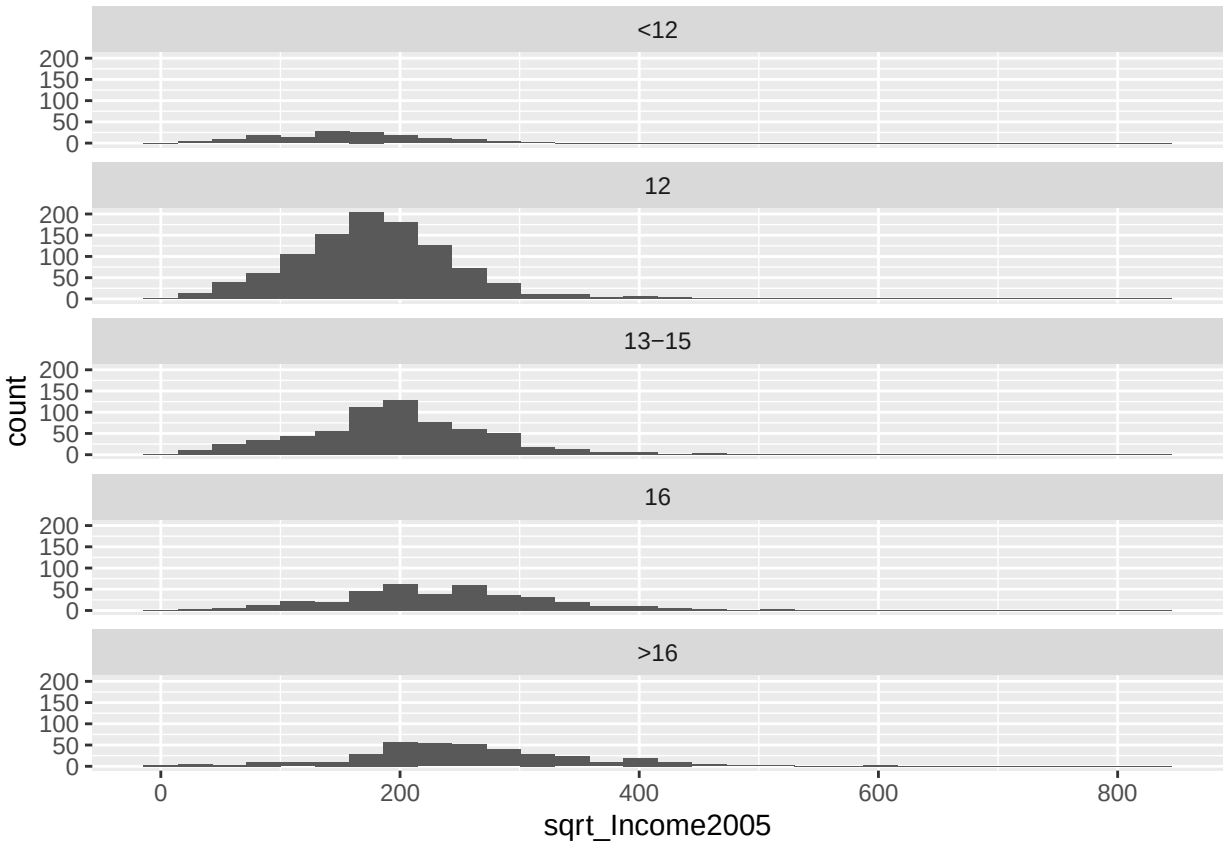
I have also added a new variable to the data frame called `sqrt_Income2005`, with the square root of each individual's income in 2005. The reason for this is that the ANOVA model asserts that the response variable follows a normal distribution within each group, but the incomes are skewed right. The 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.

```
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`.

```
ggplot(income, mapping = aes(x = sqrt_Income2005)) + geom_histogram() + facet_wrap( ~ Educ, ncol = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(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.

```
mfull <- lm(sqrt_Income2005 ~ Educ, data = income)
anova(mfull)
```

```
## Analysis of Variance Table
##
## Response: sqrt_Income2005
##             Df   Sum Sq Mean Sq F value    Pr(>F)
## Educ         4  2651561  662890  100.14 < 2.2e-16 ***
## Residuals 2579 17071986    6620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Define the following parameters:

- μ_1 = average of the square root of income in 2005 among people who have an education level of less than 12 years (less than a high school degree)
- μ_2 = average of the square root of income in 2005 among people who have an education level of 12 years (a high school degree but no college)
- μ_3 = average of the square root of income in 2005 among people who have an education level of between 13 and 15 years (some college)
- μ_4 = average of the square root of income in 2005 among people who have an education level of 16 years (a college degree)

- μ_5 = average of the square root of income in 2005 among people who have an education level of more than 16 years (an advanced degree)

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. The population mean of the square root of income is the same for all 5 education levels.

H_A : At least one of μ_1 through μ_5 is not equal to the others. The population mean of the square root of income is not the same for all 5 education levels.

The p-value for this test is less than 2.2×10^{-16}. The data provide extremely strong evidence that the population mean of the square root of income is different for people in different education levels.

(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.

```
# Using a t test
library(gmodels)
fit.contrast(mfull, "Educ", c(0, 0, 0, 1, -1))
```

```
##                       Estimate Std. Error   t value   Pr(>|t|)
## Educ c=( 0 0 0 1 -1 ) -14.01239   5.831293 -2.402965 0.01633296
## attr(,"class")
## [1] "fit_contrast"
```

```
# Using an F test
income <- income %>%
  mutate(
    educ_college_grouped = ifelse(Educ %in% c("16", ">16"), "college_group", Educ)
  )

mreduced <- lm(sqrt_Income2005 ~ educ_college_grouped, data = income)
anova(mreduced, mfull)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt_Income2005 ~ educ_college_grouped
## Model 2: sqrt_Income2005 ~ Educ
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1   2580 17110209
## 2   2579 17071986  1     38223 5.7742 0.01633 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In terms of the parameters defined in the answer to part (b), our hypotheses are:

$H_0 : \mu_4 = \mu_5$, or equivalently, $\mu_4 - \mu_5 = 0$. The population mean income among people with an undergraduate college degree is equal to the population mean income among people with a graduate degree.

$H_A : \mu_4 \neq \mu_5$. The population mean income is different among people who have an undegraduate college degree and people who have a graduate degree.

Either way you conduct the test (using a t test or an F test), the p-value is the same: 0.01633. The data provide fairly strong evidence that the mean income level is different for people with an undergraduate college degree and people with a graduate degree.

(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.

```
income <- income %>% mutate(educ_not_college_grad = ifelse(Educ %in% c("<12", "12", "13-15"), "not colle

mred <- lm(sqrt_Income2005 ~ educ_not_college_grad, data = income)
anova(mred, mfull)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt_Income2005 ~ educ_not_college_grad
## Model 2: sqrt_Income2005 ~ Educ
##   Res.Df      RSS Df Sum of Sq     F    Pr(>F)
## 1   2581 17313738
## 2   2579 17071986  2    241752 18.26 1.334e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In terms of the parameters defined in the answer to part (b), our hypotheses are:

$H_0 : \mu_1 = \mu_2 = \mu_3$. The population mean income among people without a high school degree, with a high school degree but no college education, and with some college education is the same.

H_A : At least one of μ_1, μ_2, and μ_3 is not equal to the others. The population mean income is different among people without a high school degree, with a high school degree but no college education, and with some college education.

The p-value for this test is approximately 1.33×10^{-8}. The data provide extremely strong evidence that the population mean income is not the same across people without a high school degree, with a high school degree but no college education, and with some college education is the same.