Chapter 16
Analysis of Variance

Section 16.1 and 16.2

Example 16.1 uses ANOVA to explore if the population mean GPAs are different for students who sit in the front, middle, or back seats in a classroom. The data are obtained from two datasets: UCDavis1.RData and UCDavis2.RData. To use the same data as the text to replicate the analysis, you will need to combine the data from both datasets, remove a student who gave multiple answers for seat position (this student is not apparent in the dataset), remove students who gave GPAs greater than 4.0, and ignore students who have missing data for GPA and/or seat position. The process is made more complicated because the labeling of seat position is done differently in the two datasets. The process of preparing data for analysis is called data management. Data management is often frustrating, but necessary for statistical analysis. It is a skill learned through experience and practice.

The first step is to import the data into R and remove the 78th person from the UCDavis2.RData dataset. This person responded with multiple answers for seat position. The datasets will be assigned to data frames ucdavis1 and ucdavis2. The command ucdavis2 <- ucdavis2[-78,] removes the 78 student by deleting the 78th row of the data frame, but keeping all columns (thus the blank after the comma). The R commands follow.

```
> load("C:/RData/UCDavis1.RData")
> load("C:/RData/UCDavis2.RData")
> ucdavis2 <- ucdavis2[-78,]
> ucdavis1 <- edit(ucdavis1)
> names(ucdavis1)
[1] "Sex" "TV" "computer" "Sleep" "Seat" "alcohol"
[7] "Height" "momheight" "dadheight" "exercise" "GPA" "class"
> ucdavis2 <- edit(ucdavis2)
> names(ucdavis2)
> names(ucdavis2)
[1] "Sex" "GPA" "Seat" "Alcohol" "WtFeel" "Height"
[7] "IdealHt" "momheight" "dadheight" "Hand" "Looks" "Friends"
[13] "Cheat" "Smoke"
```

Variables Seat and GPA have the same names in both ucdavis1 and ucdavis2. Using the function unique(), however, to see a list of possible values within each variable we see that ucdavis1 has levels B, M, and F while ucdavis2 has levels B, M, and F. The <NA> tells us that some students did not record any valid answer for the seat position question. The R commands and output follow.

```
> unique( ucdavis1$Seat )
[1] "B" "M" "F" "NA"
> unique( ucdavis2$Seat )
[1] "M" "F" "B" "NA"
```

Now we need to extract the GPA and Seat variables from each dataset and combine them into a single vector for seat position and a single vector for GPA. These two vectors for seat position and GPA will then be put together into a data frame for the ANOVA statistical analysis. To extract seat position and
GPA from ucdavis1 and ucdavis2 into their own vectors seat1, seat2, gpa1, and gpa2, the following commands R are used.

```r
> seat1 <- ucdavis1$Seat
> seat2 <- ucdavis2$Seat
> gpa1 <- ucdavis1$GPA
> gpa2 <- ucdavis2$GPA
```

The variables gpa1 and gpa2 can easily be combined together into a single vector using `c(gpa1, gpa2)` because they are both numeric. The list of R commands to combine the two seat vectors is shown below. Finally, we combine everything into a data frame and then remove the NA values from each column.

```r
> seats12 <- c(seat1, seat2)
> students <- data.frame(seats=seats12, gpa=c(gpa1, gpa2))
> students <- subset(students, seats != "NA") # remove from seats
> students <- na.omit(students) # remove from GPA
```

Now that the data management steps are behind you, the analysis can begin. (We will use the R function which performs ANOVA to remove the students with grade point averages greater than 4.0.) After attaching the `students` data frame, performing a boxplot is an instructive way to compare the GPA of the three groups. The following shows the input into R and the resulting boxplot.

```r
> attach(students)
> boxplot(gpa[seats=="F"], gpa[seats=="M"],
> + gpa[seats=="B"], names=c("front","middle","back"))
```
Finally the `aov()` command is used to perform ANOVA. The `aov()` function works similarly to the `lm()` function where you specify the response (y) and explanatory (x) variables using the model formula `y~x`. You can also use the `subset()` option to specify the function to use only a select part of the dataset, thus using only students who have a GPA of 4.0 or less. The `anova()` command takes the output from `aov()` and displays it in the form of an ANOVA table. The small p-value of 0.001398 suggests that there is a statistically significant difference in grade point averages between the three groups. The R commands and the `anova()` output follow.

```r
> fit <- aov( gpa ~ seats, subset=(gpa<=4), data=students )
> anova(fit)
Analysis of Variance Table
Response: gpa
              Df Sum Sq Mean Sq F value Pr(>F)
seats         2  3.994  1.997  6.6877 0.001398 **
Residuals    381 113.775  0.299
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table also shows an F-statistic of 6.6877. The p-value is the area beneath the F-distribution density curve to the right of 6.6877. The command `1-pf( 6.6877, df1=2, df2=381)` uses the cumulative distribution function for the F-distribution to find the p-value of 0.001398.

After getting statistically significant results from an ANOVA, it is of interest to find out how big the differences are between the possible pairings (front-back, middle-back, middle-front). The R command `TukeyHSD()` calculates confidence intervals that provide a 95% family confidence level for all paired differences. As explained in `help(TukeyHSD)`, the function needs the output from the `aov()` function.
as its output. Earlier we assigned the output from the `aov()` command to the variable `fit`, thus the command `TukeyHSD(fit)` results in the following output.

```
Tukey multiple comparisons of means
95% family-wise confidence level
gpa ~ seats
  Fit: aov(formula=gpa~seats, data = students, subset = (gpa <= 4))
  
  diff      lwr      upr
seats
  F-B  0.2835 0.0835  0.4835
  M-B  0.066  -0.103  0.235
  M-F -0.218 -0.380  0.055
```

The first column is the differences between the group means. The 2nd and 3rd columns are the lower and upper bounds of the 95% family-wise confidence intervals for the three possible paired differences. Looking at the output, we see that the estimated mean difference between front and back students’ GPAs is 0.2835 (front students being greater) with a confidence interval of 0.08 to 0.48 which excludes 0. This suggests the difference in GPA between front and back students is statistically significant. The Middle-Back row, however, gives a confidence interval of -0.1037 to 0.2355 which includes 0. This suggests that the difference between middle and back students is not statistically significant.

**Section 16.3: Other methods for comparing populations**

If the necessary conditions for using an ANOVA F-test are not satisfied, other methods should be tried. The Kruskal-Wallis Test is a common alternative test. Example 16.11 explores if the median number of drinks by students differs between seat positions (front, middle, back). The data are combined from variable `alcohol` found in `UCDavis1.RData` and variable `Alcohol` (note spelling) found in `UCDavis2.RData`. The R function `kruskal.test()` performs the statistical test. Look at the R output after entering the following code.

```R
> studentdrinks <- data.frame( seats=seats12, +   drinks=c(ucdavis1$alcohol, ucdavis2$Alcohol))
> studentdrinks <- subset(studentdrinks, seats!= "NA")
> kruskal.test( drinks ~ seats, data=studentdrinks )

Kruskal-Wallis rank sum test
data:  drinks by seats
Kruskal-Wallis chi-squared = 40.1815, df = 2, p-value = 1.882e-09
```

The Kruskal-Wallis statistic is 40.18 which provides a p-value < 0.001, thus we can reject the hypothesis that the median number of drinks is the same between students who sit in the front, middle, or back of a lecture hall. There is no function in R to perform Mood’s median test.