STAT502 Lab #11

In this lab we will study two-factor ANOVA with nested design.

1. (*from Ch.26 example*) There are three schools of interest (Factor A), each in a different city, and each school has exactly two resident instructors (Factor B). The goal is to study how teaching performance depends on the schools and instructors. At the end of the study, two teaching scores are collected from each instructor at each city. The data can be read into R with the following commands:

```
scores = c(25,29,14,11,11,6,22,18,17,20,5,2)
A = as.factor(c(rep('Atlanta',4),rep('Chicago',4),rep('SanFran',4)))
B = as.factor(c(1,1,2,2,3,3,4,4,5,5,6,6))
B2 = as.factor(c(1,1,2,2,1,1,2,2,1,1,2,2))
```

   Note that the "B2" factor is an alternate way to label the instructors. It reflects the fact that there are two instructors per city but not that every instructor is unique—that they are nested within a particular city. This is something to be aware of when working with nested data; it could be coded either way.

   (a) The difference in the instructor codings is also apparent when constructing boxplots. Notice the difference between the following commands. In each case, what values are included in each box? Why is the second one inappropriate?

```
boxplot(scores ~ B)
boxplot(scores ~ B2)
```

   (b) Using the first appropriate plot above, comment on the effect of instructor. Specifically, which instructors are being compared?

   (c) We can also use `boxplot(scores  A)` to plot the data for the cities. This is ok since all observations in each box come from the same city. Comment on the effect of city.

   (d) The two-way nested ANOVA model with B nested within A (both fixed) is given by

$$Y_{ijk} = \mu_{\cdot} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

   where $\sum_i \alpha_i = \sum_j \beta_{j(i)} = 0$, and $\epsilon_{ijk} \sim N(0, \sigma^2)$. It can be fit in R with the commands

```
fit = lm(scores ~ A/B)
anova(fit)
```

   The symbol '/' is used to indicate Factor B is nested within Factor A. We could also have used `lm(scores~A/B2)` or even `lm(scores~A+B)` since B is coded "correctly" with separate labels for each instructor. The one combination that doesn't work is `lm(scores~A+B2)`. In this case, R has no way of knowing B is nested.

   (e) Report the ANOVA table for this model. When compared with the crossed ANOVA model (if each of two instructors taught at all three cities), the difference that the sum of squares normally separated for B and AB interaction are effectively combined together. To see this, use `anova(lm(scores~A+B2+A*B2))` to fit the (incorrect!) ANOVA model that assumes crossed factors. Compare the degrees of freedom and sum of squares with the correct ANOVA table.

(f) What error term is used for each of the tests? What effects are significant? Report the test statistics and p-values.

(g) Use the fact that

$$\frac{(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{i'\cdot\cdot}) - (\alpha_i - \alpha_{i'})}{\sqrt{MSE(\frac{2}{bn})}} \sim t_{ab(n-1)}$$

to construct 90% confidence intervals for the pairwise differences $(\alpha_i - \alpha_{i'})$ among cities. Use the Tukey procedure to adjust for multiplicity. Which are significantly different?

(h) Use the fact that

$$\frac{(\overline{Y}_{ij\cdot} - \overline{Y}_{ij'\cdot}) - (\beta_{j(i)} - \beta_{j'(i)})}{\sqrt{MSE(\frac{2}{n})}} \sim t_{ab(n-1)}$$

to construct a 90% confidence interval for the difference $(\beta_{1(1)} - \beta_{2(1)})$ between instructors for the first city. Repeat this for the other two cities. Use a Bonferroni adjustment for multiplicity. In which cities are instructors significantly different?

2. (*adapted from Exercise 26.4*) The goal in this problem is to study the effects of machine (Factor A) and operator (Factor B) on the output in a bottling plant. Three machines were used, and twelve operators were employed. Four operators were assigned to a machine. Data on the number of cases produced by each machine and operator were collected and are available in the data set “bottling.txt”.

```
data = read.table('bottling.txt', header=T)
```

The first column is the number of cases, the second column is the machine (1,2,3), and the third column is the operator (1,2,3,4).

(a) Which factor is nested within which factor? It may help to sketch a diagram.

(b) Assuming both factors are fixed (they are the only levels of interest), express the model for Y_{ijk}, the number of cases for the kth day and the jth operator using the ith machine. Also, state the ranges for i, j, and k.

(c) Fit the model to the data, and report the ANOVA table. Which effects are significant?

(d) Use Tukey's procedure to conduct pairwise comparisons for the mean cases of the three machines. What conclusions can be made?