STAT 461: Lab 8 - ANOVA For Nested Designs

1 Review: Factorial (or Crossed) Experimental Designs

Consider an experiment in which there are two factors (types of treatments), which we will call "Factor A" and "Factor B". If every possible combination of (1) the levels of Factor A and (2) the levels of Factor B are applied to experimental units, then we say that we have a **Factorial Treatment Design** or that **Factor A and Factor B are Crossed** with each other.

1.1 The Two-Way Complete ANOVA Model

The two-way complete ANOVA model for a factorial design is

$$Y_{ijt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijt}, \quad \epsilon_{ijt} \overset{iid}{\sim} N(0, \sigma^2)$$

$$i = 1, 2, \ldots, a \qquad j = 1, 2, \ldots, b \qquad t = 1, 2, \ldots, r_{ij}$$

Where the first index i indexes the levels of Factor A, with

$$i = 1, 2, \ldots, a = \text{ the number of treatment levels of Factor A}$$

and the second index j indexes the levels of Factor B, with

$$j = 1, 2, \ldots b = \text{ the number of treatment levels of Factor B.}$$

The **main effects** are α_i, which encodes the effect of the i-th level of Factor A on the mean response, and β_j, which is the effect of the j-th level of Factor B on the mean response.

The **interaction effects** are $(\alpha\beta)_{ij}$ and encode the additional effect on the mean response of the joint combination of level i of Factor A and level j of Factor B.

2 Nested Designs

A 2-way **nested** design is one in which each level of Factor B can only be combined with one level of Factor A. The model for a nested design in which Factor B is nested within Factor A is

$$Y_{ijt} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijt}, \quad \epsilon_{ijt} \overset{iid}{\sim} N(0, \sigma^2)$$

$$i = 1, 2, \ldots, a \qquad j = 1, 2, \ldots, b \qquad t = 1, 2, \ldots, r_{ij}$$

Where the first index i indexes the levels of Factor A, with

$$i = 1, 2, \ldots, a = \text{ the number of treatment levels of Factor A}$$

and the second index j indexes the levels of Factor B, with

$$j = 1, 2, \dots b = \text{ the number of treatment levels of Factor B.}$$

The only **main effect** is α_i, which encodes the effect of the i-th level of Factor A on the mean response.

The effect of the j-th level of Factor B is only present in the **nested effect** $\beta_{j(i)}$, as each level of B is only paired with a single level of Factor A. This effect encodes the additional effect on the mean response of the joint combination of level i of Factor A and level j of Factor B.

2.1 Estimable Functions

Recall that a function of parameters is estimable if it is a linear combination of the treatment means $E(Y_{ijt})$. From the one-way ANOVA representation in Section 2.1, we can see that estimable functions of the parameters in the two-way complete model can be written as

$$\sum_i \sum_j c_{ij}(\widehat{\mu} + \widehat{\tau}_{ij}) = \sum_i \sum_j c_{ij}(\widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta_{j(i)}})$$

for real numbers $\{c_{ij}\}$.

3 Hypothesis Tests

In the factorial model, we had an order to our hypothesis tests. We first tested for the significance of the interaction term. If it was significant, we looked at pairwise interactions. If it is not, we can compare the levels of the main effects A and B.

In a nested model, having the levels of B nested within A makes it possible to do the following tests in all cases.

1. A test for the overall effect of Factor A. The null hypothesis is that the mean effect of each level of A, **averaged over all nested levels of B** is the same.

$$H_0 : \alpha_i + \overline{\beta}_{\cdot(i)} = \alpha_i + \overline{\beta}_{\cdot(i)} = \cdots = \alpha_i + \overline{\beta}_{\cdot(a)}$$

Note that if all the effects of B are equal to zero ($\beta_{j(i)} = 0$ for all i, j), then this reduces to our usual null hypothesis for testing for differences in the levels of Factor A.

The test statistic used is

$$\frac{SSA/(a-1)}{SSE/(n-ab)} \sim F_{a-1,n-ab}$$

where the distribution is under the null.

2. For the nested Factor B, we can test whether or not the level of B affects the mean response associated with the level of A. The joint null hypothesis is

$$H_0 : \left\{ \begin{array}{c} \beta_{1(1)} = \beta_{2(1)} = \cdots = \beta_{b(1)} \\ \beta_{1(2)} = \beta_{2(2)} = \cdots = \beta_{b(2)} \\ \vdots \\ \beta_{1(a)} = \beta_{2(a)} = \cdots = \beta_{b(a)} \end{array} \right\}$$

Under this null hypothesis, the test statistic used is

$$\frac{SSB(A)/b(a-1)}{SSE/(n-ab)} \sim F_{b(a-1),n-ab}.$$

3. We can also test for pairwise differences between all treatment combinations:

$$H_0 : \tau_{11} - \tau_{12} = 0$$
$$H_0 : \tau_{11} - \tau_{13} = 0$$
$$H_0 : \tau_{21} - \tau_{12} = 0$$
$$\dots$$

and so on.

4 Example: Training

A national firm has three training schools in three cities (Atlanta, Chicago, and San Fransisco). Each school has two instructors, who only teach at one school. Each instructor teaches two classes, and the learning achievement for each class is measured with a test.

It this situation,

*Y_{ijt} is the learning achievement of the t-th class taught by instructor j at school i.

*The instructors are **nested** within school - each instructor only teaches at one school.

The nested ANOVA model for this data is

$$Y_{ijt} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijt}, \quad \epsilon_{ijt} \overset{iid}{\sim} N(0, \sigma^2)$$

$$i = A, C, SF \qquad j = 1, 2 \qquad t = 1, 2$$

The following code reads the data into R and fits this nested model. The ANOVA table for this analysis is

```r
School<-c(rep("Atlanta", 4), rep("Chicago", 4), rep("SanFran", 4))
Instructor<-c(rep(c(1,1,2,2), 3))
response<-c(25,29,14,11,11,6,22,18,17,20,5,2)
df<-data.frame(School=as.factor(School), Instructor=as.factor(Instructor), response=response)


options(contrasts = c("contr.sum", "contr.poly"))
model1<-aov(response~School/Instructor, data=df)
model1<-aov(response~School+Instructor:School, data=df)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: response
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## School             2  156.5   78.25  11.179 0.009473 **
## School:Instructor  3  567.5  189.17  27.024 0.000697 ***
## Residuals          6   42.0    7.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that we specify the model using the following line

`response~School/Instructor`

This can also be done by

`response~School+School:Instructor`

where the `School` term specifies the main effect of School and the `School:Instructor` term tells R that `Instructor` is nested within `School`.

The residuals for this model fit show that the assumption of constant error variance is met, and that normality is at least marginally satisfied.

The `School:Instructor` row in the ANOVA table has the test for the effect of instructor within schools. As this is significant, we reject the null hypothesis and conclude that classes taught by different instructors at the same school have different mean achievement levels.

The `School` row in the ANOVA table has the test for the effect of school. As this is significant, we conclude that different schools have different mean achievement levels.

We can then look at pairwise differences between the levels of Factor A (school). We can also look at pairwise differences between instructor/school combinations (instructors nested within schools). The following code tests for all these pairwise differences.

Pairwise differences between the levels of Factor A (school)

The Tukey groupings are

```
library(lsmeans)
library(multcompView)

lsmSchool=lsmeans(model1, ~ School )

cld(lsmSchool, alpha=0.05)
```

```
##  School  lsmean       SE df lower.CL upper.CL .group
##  SanFran  11.00 1.322876  6  7.76304 14.23696  1
##  Chicago  14.25 1.322876  6 11.01304 17.48696  12
##  Atlanta  19.75 1.322876  6 16.51304 22.98696   2
##
## Results are averaged over the levels of: Instructor
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 3 estimates
## significance level used: alpha = 0.05
```

Pairwise differences between instructor/school combinations (instructors nested within schools)

The Tukey groupings are

```
lsmSchoolInstr=lsmeans(model1, ~ School:Instructor )

cld(lsmSchoolInstr, alpha=0.05)
```

```
##  Instructor School  lsmean       SE df  lower.CL  upper.CL .group
##  2          SanFran    3.5 1.870829  6 -1.077753  8.077753  1
##  1          Chicago    8.5 1.870829  6  3.922247 13.077753  12
```

```
## 2          Atlanta   12.5 1.870829  6  7.922247 17.077753  123
## 1          SanFran   18.5 1.870829  6 13.922247 23.077753   234
## 2          Chicago   20.0 1.870829  6 15.422247 24.577753    34
## 1          Atlanta   27.0 1.870829  6 22.422247 31.577753     4
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 6 estimates
## significance level used: alpha = 0.05
```

The results for school show that Atlanta classes have significantly higher mean achievement than do classes taught in San Francisco.

The results for instructor show, for example, that mean achievement in classes taught by Instructor 2 in Chicago is significantly higher than mean achievement in classes taught by Instructor 1 in Chicago. There are many other comparisons that could be made as well.

Homework Assignment

1. **Widgets**. Download the "widgets.csv" code from Canvas. Input data into R using the following code.

```
widgets=read.table("widgets.txt",header=TRUE)
Batch=as.factor(widgets$Batch)
Supplier=widgets$Supplier
WidgetSize=widgets$WidgetSize
```

Suppose that you have a business that uses widgets as raw material for a product that you manufacture and sell. The widgets are currently purchased from several different suppliers. You notice that your machines are constantly breaking down, and find out that this is due to excessive variability in the sizes of the widgets. If you can make sure that your widgets are more uniform in size, then you can adjust your machines to allow for this.

The widget variability could be due to any of three reasons: (1) differences between suppliers, in which case you could restrict to one or two suppliers at a time, or else tell each of your factories to purchases widgets only from the the supplier that is closest, (2) differences between average sizes for different batches of widgets from the same same supplier. In that case, you can test one or two widgets from each box and return boxes with widgets whose average size is too large or too small, or (3) variability within each batch of widgets, in which case there may be no simple solution other either buying (or designing) more robust machines or else negotiating with your suppliers to provide more uniform widgets.

As a way to determine the source of the machine-damaging size variation in the widgets, you select 3 suppliers and order 4 boxes of widgets from each supplier at different times. You select 4 widgets from each box for testing, for a total of 3 x 4 x 4 = 48 widgets.

(a) Write out the 2-way nested model for this experiment.

(b) Explain why one factor is a "nested" factor.

(c) Fit the model using R and examine the residuals. Transform the response if needed to address any problems with normality or constant error variance. If you transform the response, clearly show the residuals from the un-transformed response, and your best transformation, and describe why you chose the transformation you did.

(d) Identify sums of squares that could be used to partition the variation in widget size into (1) variation due to differences between suppliers, (2) differences between averae sizes for different batches of widgets from the same supplier, and (3) variability within each batch of widgets. Which of these sums of squares is the greatest for this dataset?

(e) Interpret the results of the hypothesis tests you obtained using R in the context of this experiment.

2. For the following descriptions of experiments, write out an appropriate model for the experiment. Clearly show which factors are nested and which are crossed. It may be helpful (but is not required) to draw a diagram of the experiment.

(a) An experiment is conducted in which 30 runners (15 men and 15 women) were randomly assigned to one of three physical training groups (endurance training, strength training, and cross training). The randomization is conducted so that 5 men and 5 women are assigned to each of the three training groups. The race times of the runners were recorded both before and after training for one month. The response variable of interest is the difference between before and after race times.

(b) An experiment is conducted to test how much time students from different majors spend studying. Three students are randomly selected from all students majoring in Statistics. Three seperate students are randomly selected from all students majoring in Humanites. Each student recorded how much time they spent studying on three different days.

(c) An engineer wants to study variability in the strength of glass cathode supports made on production machines in the manufacturing plant. There are five production machines in the plant, and each machine has 4 components called ‘heads’ which produce the glass cathode supports. The engineer took 4 glass supports made from each head and tested their strength. Data collection of the 5x4x4 = 80 measurements was completely randomized.