STAT 461 Lab 5 - ANOVA in R

1. ANOVA Model and Important Hypotheses

Recall from lectures that the one-way ANOVA model with effects coding is written

$$Y_{it} = \mu + \tau_i + \epsilon_{it}, \quad i = 1, 2, \ldots, v \quad t = 1, 2, \ldots, r_i$$

$$\epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2)$$

The v treatments are indexed by i and the number of replicates receiving the i-th treatment is r_i. The sample mean for the i-th treatment is

$$\bar{Y_{i\cdot}} = \frac{1}{r_i} \sum_{t=1}^{r_i} Y_{it}$$

and the grand mean is

$$\bar{Y_{\cdot\cdot}} = \frac{1}{n} \sum_{i=1}^{v} \sum_{t=1}^{r_i} Y_{it}$$

1.1 Testing for Equality of All Treatment Effects

The first hypothesis tested in an experiment is that there is any difference in mean response between the treatments. This is expressed as the null hypothesis:

$$H_0 : \tau_1 = \tau_2 = ... = \tau_v$$

1.2 Testing for Pairwise Differences Between Treatments

If a test of the above hypothesis shows that there are some differences in mean response between treatments, a common next step is to examine where those differences are. If the working hypothesis is that there is a difference between treatment means for treatment 1 and treatment 2, then we could test this by considering

$$H_0 : \tau_1 - \tau_2 = 0$$

We could similarly test all possible pairwise differences, one at a time:

$$H_0 : \tau_1 - \tau_3 = 0$$

$$H_0 : \tau_2 - \tau_3 = 0$$

and so on.

2. The ANOVA Table

The ANOVA table will provide us with a bookeeping tool to compute the test statistic for the equality of all treatment means.

$$H_0 : \tau_1 = \tau_2 = ... = \tau_v$$

2.1 Important Sums of Squares

We build a test statistic for testing this null hypothesis based on sums of squares.

- the **Total Sum of Squares** quantifies the variation in the data $\{Y_{it}\}$ around the grand mean $\bar{Y_{..}}$

$$SSTOT = \sum_{i=1}^{v} \sum_{t=1}^{r_i} (Y_{it} - \bar{Y_{..}})^2$$

If σ^2 is known, then we have

$$\frac{SSTOT}{\sigma^2} \sim \chi^2_{n-1}$$

We can partition the SSTOT into the following two sums of squares:

- the **Treatment Sum of Squares** quantifies the variation in the treatment means $\{\bar{Y_{i.}}\}$ around the grand mean $\bar{Y_{..}}$

$$SST = \sum_{i=1}^{v} r_i (\bar{Y_{i.}} - \bar{Y_{..}})^2$$

If σ^2 is known, then we have

$$\frac{SST}{\sigma^2} \sim \chi^2_{v-1}$$

* The **Sum of Squared Errors** quantifies the variation in the data $\{Y_{it}\}$ around the treatment means $\{\bar{Y_{i.}}\}$

$$SSE = \sum_{i=1}^{v} \sum_{t=1}^{r_i} (Y_{it} - \bar{Y_{i.}})^2$$

If σ^2 is known, then we have

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-v}$$

2.2 The ANOVA F-Test Statistic

As discussed in class, a test statistic T^* for $H_0 : \tau_1 = \cdots = \tau_v$ is

$$T^* = \frac{SST/(v-1)}{SSE/(n-v)} \sim F_{v-1,n-v}$$

where $T^* \sim F_{v-1,n-v}$ under H_0. We can thus reject H_0 if T^* is larger than could be expected by random chance. We can compute the p-value

$$p = P(T > T^*) \text{ given that } T \sim F_{v-1,n-v}$$

and reject H_0 if $p < 0.05$. The p-value can be computed by looking at standard tables, or by using online calculators such as the one at http://graphpad.com/quickcalcs/PValue1.cfm Note that we are conducting a one-sided test. That is, we only reject H_0 if T^* is BIGGER than expected by random chance under the null.

2.3 The ANOVA Table

We can encode the calculations needed to construct the test statistic T^* in a table. The columns denote different quantities needed, and the rows denote the different sources of variation.

Source	Deg. of Freedom	Sum of Squares	Mean Square	Ratio
Treatment	v-1	SST	$SST/(v-1)$	$\frac{SST/(v-1)}{SSE/(n-v)}$
Error	$n-v$	SSE	$SSE/(n-v)$	
Total	$n-1$	$SSTOT$		

This table is just a bookeeping tool to help you do the calculations needed to arrive at the test statistic. An important note is that in the first two columns, the "Treatment" and "Error" rows sum up to the "Total" row. Thus you **only need to calculate two of the sums of squares**. Often it is easier to calculate SST and SSTOT than it is to calculate SSE.

3 Example: Catching Flies

As an example of a hypothesis test, consider the following experiment. The aim was to test the maxim: "You can catch more flies with honey than with vinegar". The intended meaning of this statement is that being nice to others will benefit you more than being mean. We will not be testing the effects of being nice, but instead testing how attracted flies are to honey, vinegar, or water.

Eight bowls were laid out on a table in two rows. At random, each bowl was filled with either honey, vinegar, or water. The randomization was conducted so that three of the bowls were filled with honey, three with vinegar, and two with water. The bowls were watched for 10 minutes, and observers recorded each time a fly landed either on a bowl or the liquid in the bowl. At the end of 10 minutes, the total number of flies landing on each bowl was recorded:

Honey	Vinegar	Water
2	4	2
1	3	3
2	4	

We will denote the three treatments with the subscripts: H = Honey V = Vinegar W = Water

I typically fill out the ANOVA table from left to right:

1. Fill in the degrees of freedom column. This comes just from the study design.
2. Calculate all sample treatment means $\{Y_{i.}, i = 1, 2, ..., v\}$ and the grand mean $Y_{..}$.
3. Use the $\{Y_{i.}\}$ and $Y_{..}$ to compute SST SSTOT in the "Sum of Squares" column.
4. Fill in the Error box i the "Sum of Squares" column by computing SSE=SSTOT-SST.

5. Compute the mean squares by dividing SST and SSE by their respective degrees of freedom.
6. Calculate the test statistic in the "Ratio" column by dividing the two mean sums of squares.

The following page is a worksheet to help you fill out the ANOVA table.

The ANOVA Table

Source	Deg. of Freedom	Sum of Squares	Mean Square	Ratio
Treatment				
Error				
Total				

1. Degrees of Freedom Treatment = v-1, Error = $n - v$, Total = n-1

2. Treatment Means and Grand Mean

$$\bar{Y_{i.}} = \frac{1}{r_i}\sum_{t=1}^{r_i} Y_{it} \quad , \quad \bar{Y_{..}} = \frac{1}{n}\sum_{i=1}^{v}\sum_{t=1}^{r_i} Y_{it}$$

3. Sums of Squares

$$SSTOT = \sum_{i=1}^{v}\sum_{t=1}^{r_i}(Y_{it} - \bar{Y_{..}})^2 \qquad SST = \sum_{i=1}^{v} r_i(\bar{Y_{i.}} - \bar{Y_{..}})^2 \qquad SSE = \sum_{i=1}^{v}\sum_{t=1}^{r_i}(Y_{it} - \bar{Y_{i.}})^2$$

4 ANOVA in R

The functions aov() and anova() calculates the ANOVA table, the test statistic T^*, and the p-value of the test statistic automatically. The following code reads in the data into R. It treats liquid as a factor-type variable.

```
liquid<-c(rep("Honey", 3), rep("Vinegar", 3), rep("Water", 2))
numFlies<-c(2,1,3,5,4,5,1,3)
data<-data.frame(liquid=as.factor(liquid), numFlies=numFlies)
data
```

```
##    liquid numFlies
## 1   Honey        2
## 2   Honey        1
## 3   Honey        3
## 4 Vinegar        5
## 5 Vinegar        4
## 6 Vinegar        5
## 7   Water        1
## 8   Water        3
```

The following code fits the One-Way ANOVA model to test the null hypothesis of no difference in treatment means.

```
library(knitr)

model1<-aov(numFlies~liquid, data = data)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: numFlies
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## liquid     2 13.3333  6.6667  7.1429 0.03422 *
## Residuals  5  4.6667  0.9333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To get better looking output use the function kable() in the library "knitr".

```
library(knitr)
kable(anova(model1), format = "markdown")
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
liquid	2	13.333333	6.6666667	7.142857	0.0342244
Residuals	5	4.666667	0.9333333	NA	NA

From this we see that the test statistic is $T^* = 7.14$, and the p-value is 0.0342. This indicates that under the null hypothesis that there is no difference in treatment means, that we will observe test statistics greater than 7.14 only about 3% of the time. As this is less than 0.05 (our usual level for a hypothesis test), we reject the null hypothesis and conclude the following:

Result: The liquid in a bowl (water, vinegar, or honey) significantly affects the mean number of flies landing on the bowl. Note that this doesn't say anything about which treatments are different from each other, but only that there are some differences.

Alternatively Construction of decision rule by using F distribution

Find the 95th percentile of the F distribution with (2, 5) degrees of freedom.

Apply the quantile function *qf* of the F distribution against the decimal value 0.95.

```
qf(0.95, df1 = 2, df2 = 5)
```

```
## [1] 5.786135
```

Find p value:

```
pf(F value, df1, df2, lower.tail = F)
```

Homework Assignment

1. Show that $SSTOT/\sigma^2 \sim \chi^2_{n-1}$ by using an argument similar to that used in the lectures to find distributions for SSE/σ^2 and SST/σ^2.
2. A test statistic that could be used to test for a significant pairwise differences between the i-th and j-th treatment is

$$D^*_{ij} = \frac{(\bar{Y_{i.}} - \bar{Y_{j.}})^2}{SSE/(n-v)}$$

The null hypothesis is $H_0 : \tau_i = \tau_j$. Find the distribution of D^*_{ij} under the null hypothesis.

(For your information, R uses the square root of this statistic to get the p-values reported in the lsmeans output)

3. Recall the soap experiment from Homework 1. Look back at Homework 1 for an explanation of the experiment. The data are the weight lost over 24 hours by different types of soap.

Cube	**Regular**	**Deodorant**	**Moisturizing**
1	-0.30	2.63	1.86
2	-0.10	2.61	2.03
3	-0.14	2.41	2.26
4	0.40	3.15	1.82

(a) Construct the ANOVA table for this experiment by hand. Show the calculations needed to construct the quantities in the ANOVA table.

(b) Test the null hypothesis that there is no difference in mean weight lost between different soap types. Report the test statistic, the p-value of the statistic under the null hypothesis, and interpret the result of the test. You may use R or an online p-value calculator (like http://graphpad.com/quickcalcs/PValue1.cfm) to compute the p-value. Provide the R code used, with output, or clearly describe the online p-value calculator you used.