Lab 10: ANOVA

For this lab, it will be helpful to have a copy of the knitted version of this document to answer the questions as much of it is written using mathematical notation that may be difficult to read when the document is not knitted.

Lab Goals

The purpose of this lab is to develop a deeper understanding of an ANOVA hypothesis test. In this lab you will:

1) Calculate an ANOVA statistic "by hand."
2) Learn how to perform an ANOVA hypothesis test in R.
3) Learn how to check the assumptions associated with an ANOVA hypothesis test.

ANOVA

An ANOVA hypothesis test is used to test the

null hypothesis, $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ versus the

alternative hypothesis, $H_a :$ At least one $\mu_i, i = 1, .., k$ is different from the others.

ANOVA Test Statistic

Under H_0 the test statistic $F^* =$ MSB/MSE has an $F_{df1,df2}$ distribution where:

1) MSB = SSB/df1 is the mean square between the k groups and
    a) SSB = $\sum_{i=1}^{k} n_i \ (\bar{Y}_{i.} - \bar{Y}_{..})^2$
    b) df1 = k-1 = between groups degrees of freedom
2) MSE = SSE/df2 is the mean squared error
    a) SSE = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$
    b) df2 = n_{tot} - k = error degrees of freedom where n_{tot} corresponds to the total sample size over all k groups

Note the Following Relationships

1. TSS = SSB + SSE where TSS = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$
2. Total degrees of freedom = n_{tot}-1 = df1 + df2, where n_{tot} is the total sample size over all k groups

Problem 1

A study was designed to assess the effectiveness of three fertilizers, A, B and C, for improving the yield of a certain variety of wheat. There were a total of 12 plots, with each fertilizer applied to 4 plots of wheat. The yield (in bushels) was recorded. The research question of interest is whether there is a difference in the mean wheat yield between the 3 fertilizers. Assume the ANOVA assumptions are satisfied for this study.

Let y_{ij} denote the yield of wheat from plot j with fertilizer i for i = 1,2,3 and j = 1,2,3,4. The observed data are contained in the table below.

Fertilizer				
A	$y_{11} = 3$	$y_{12} = 4$	$y_{13} = 5$	$y_{14} = 4$
B	$y_{21} = 2$	$y_{22} = 4$	$y_{23} = 3$	$y_{24} = 3$
C	$y_{31} = 4$	$y_{32} = 6$	$y_{33} = 5$	$y_{34} = 5$

Here we will calculate the ANOVA statistic and perform the test "by hand." Include a code chunk for each calculation.

a) Calculate the mean yield for each fertilizer, $\bar{y}_{i.}$ for i = 1,2,3.

i. $\bar{y}_{1.} = 4$
ii. $\bar{y}_{2.} = 3$
iii. $\bar{y}_{3.} = 5$

```
A_mean=mean(c(3,4,5,4))
B_mean=mean(c(2,4,3,3))
C_mean=mean(c(4,6,5,5))

A_mean
```

```
## [1] 4
```

```
B_mean
```

```
## [1] 3
```

```
C_mean
```

```
## [1] 5
```

b) Calculate the grand mean, $\bar{y}_{..}$.

```
GM=mean(c(3,4,5,4,2,4,3,3,4,6,5,5))
GM
```

```
## [1] 4
```

c) Calculate the tss (total sum of squares) = $\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$.

```
sst=sum((c(3,4,5,4,2,4,3,3,4,6,5,5)-GM)^2)
sst
```

```
## [1] 14
```

d) Calculate the ssb (sum of squares between groups)= $\sum_{i=1}^{K} n_i \ (\bar{y}_{i.} - \bar{y}_{..})^2$.

```
ssb=4*sum((c(A_mean,B_mean,C_mean)-GM)^2)
ssb
```

```
## [1] 8
```

e) Using (c) and (d), calculate sse (error sum of squares).

```
sse = sst-ssb
sse
```

```
## [1] 6
```

f) Find the between groups degrees of freedom (df1).

```
df1=3-1
df1
```

```
## [1] 2
```

g) Find the total degrees of freedom.

```
12-1
```

```
## [1] 11
```

h) Find the error degrees of freedom (df2).

```
df2=12-3
df2
```

```
## [1] 9
```

i) Compute the msb (mean square between groups).

```
msb=ssb/df1
msb
```

```
## [1] 4
```

j) Compute the mse (mean squared error).

```
mse=sse/df2
mse
```

```
## [1] 0.6666667
```

k) Determine the realization of the test statistic, $F^* = \text{msb}/\text{mse}$.

```
Fstar=msb/mse
Fstar
```

```
## [1] 6
```

l) Determine the rejection region of this test using a significance level of 0.05, RR: $F^* > F_{0.05,df1,df2}$. Based on the rejection region, should we reject H_0?

```
qf((1-0.05),df1,df2)
```

```
## [1] 4.256495
```

Yes, since $6 > 4.256$*, we should reject* H_0.

m) Determine the p-value of this test, $P(F_{df1,df2} > F^*)$. Include the formula for the p-value.

$P(F_{2,9} > 6)$=

```
1-pf(Fstar,df1,df2)
```

```
## [1] 0.02208536
```

n) Based on the p-value, should we reject H_0 at the 0.05 significance level? State your conclusion in the context of the problem.

Yes, since $0.022 < 0.05$*, we should reject* H_0*. We are 95% confident that mean yield is not the same for all three fertilizer treatments.*

Problem 2

A new bakery is opening in Ithaca, and the head baker would like to do a little market research geared towards determining what types of cookies are most popular in Ithaca. Ninety Ithacans are randomly sampled to participate in this study. They are each randomly assigned one of 6 different types of cookies to sample (15 per type of cookie). The 6 groups of cookies are: chewy chocolate chip, crispy chocolate chip, chewy sugar, crispy sugar, chewy oatmeal and crispy oatmeal. After each participant samples his cookie, he rates it on a scale of 1-100 for tastiness where higher scores correspond to more tasty cookies. The *Tastiness.csv* file contains the data for this study.

Assume:

μ_1 = mean tastiness of chewy chocolate chip cookies

μ_2 = mean tastiness of crispy chocolate chip cookies

μ_3 = mean tastiness of chewy sugar cookies

μ_4 = mean tastiness of crispy sugar cookies

μ_5 = mean tastiness of chewy oatmeal cookies

μ_6 = mean tastiness of crispy oatmeal cookies

a) Read the data into this lab document. Additionally, it is a good idea to get into the habit of making sure R is interpreting your categorical variables correctly. The code, `is.factor(Tastiness$cookie)` will let you check that `cookie` is a categorical variable. (If it is not, you can use `Tastiness$cookie <- as.factor(Tastiness$cookie)` to specify to R that it should be treated as a categorical variable.)

```
Tastiness <- read.csv("Tastiness.csv")
is.factor(Tastiness$cookie) # don't need any further conversion
```

```
## [1] TRUE
```

b) State the null and alternative hypotheses of an ANOVA analysis of these data.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

H_a: *At least one of the means is not the same as the others*

c) The ANOVA table can be generated in two steps. The following code will first run a linear model named `tasty.lm` with `rating` as the response and `cookie` as the treatment. The `anova()` function with `tasty.lm` as its argument will produce the ANOVA table for this analysis.

```
tasty.lm = lm(rating ~ cookie, data = Tastiness)
anova(tasty.lm)
```

```
## Analysis of Variance Table
##
## Response: rating
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cookie     5 1324.6 264.927  13.447 1.245e-09 ***
## Residuals 84 1654.9  19.701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) What is the p-value for the ANOVA test? State the p-value here with the correct formula.

$P(F_{5,84} > 13.45) = 1.2E-09$

e) Based on (d), state the conclusion of the ANOVA test if the significance level of the test is 0.01.

Since 1.2E-09 < 0.01, we reject the null hypothesis that average tastiness is the same for all 6 cookie varieties.

ANOVA Assumptions

In order to justify using an ANOVA hypothesis test the following assumptions must be satisfied:

1) All observations, Y_{ij}, $i = 1...k$ and $j = 1...n_k$ are mutually independent.
2) $Y_{ij} \sim N(\mu_i, \sigma)$ for each $i = 1...k$ and $j = 1...n_k$. This implies the following:
    i) All n_i observations associated with population i are drawn from a normal distribution with mean μ_i.

ii) The variance of the normal distribution is the same for all populations.

These assumptions are often restated in the following way:

1) For an observational study, independent SRSs are collected from k different populations. In an experimental study the equivalent assumption is that a SRS of sample units are each randomly assigned to one of k treatments.

2) Observations can be modeled by $Y_{ij} = \mu_i + \epsilon_{ij}$ or equivalently $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ where

   a) Y_{ij} is the jth observation from population i

   b) $\mu_i = \mu + \alpha_i$ = mean for population i, where

      1. μ = overall mean
      2. α_i = fixed effect of population i

   c) ϵ_{ij} = residual for the jth observation from population i (i.e. the difference between the j-th observation for population i and the mean of population i)

   d) ϵ_{ij} $i = 1...k$, $j = 1...n_k$ are iid $N(0, \sigma)$

Checking the ANOVA Assumptions

1) **Independence** If the populations are sampled from as stated above, this assumption is considered satisfied.

2) **Normality of the Residuals** The population of residuals is assumed to be normally distributed. This can be assessed by creating a Q-Q plot of the residuals.

3) **Equal Variance of the Distributions of All k Populations**

   a) Boxplots can be created to get a general sense of whether the sample variances are close enough to indicate the k population variances are equal.

   b) *Levene's test* can be performed to formally check whether it seems reasonable to assume the population variances are equal. The null hypothesis of the Levene test is that all population variances are equal. The Levene test can be thought of as an ANOVA performed on the absolute differences between each observation and its estimated group mean (a measure of variability). The numerator of the test statistic estimates the variability between the average absolute differences over all groups. The denominator estimates the variability of absolute differences within each group. If these estimates are similar, the Levene test will fail to reject the null hypothesis that all population variances are equal.

Problem 3

The `lm()` function in R also produces output pertaining to the estimated linear model. In particular, it provides the estimated residuals $\hat{\epsilon}_{ij}$, $i = 1...k$ and $j = 1...n_i$ associated with each observation y_{ij}. Use the following code to store the estimated residuals for the cookie study described in Problem 2.

```
tasty.res=tasty.lm$residuals  #Stores Estimated Residuals
```

We can use a Q-Q plot to determine whether it seems plausible that the residuals are normally distributed.

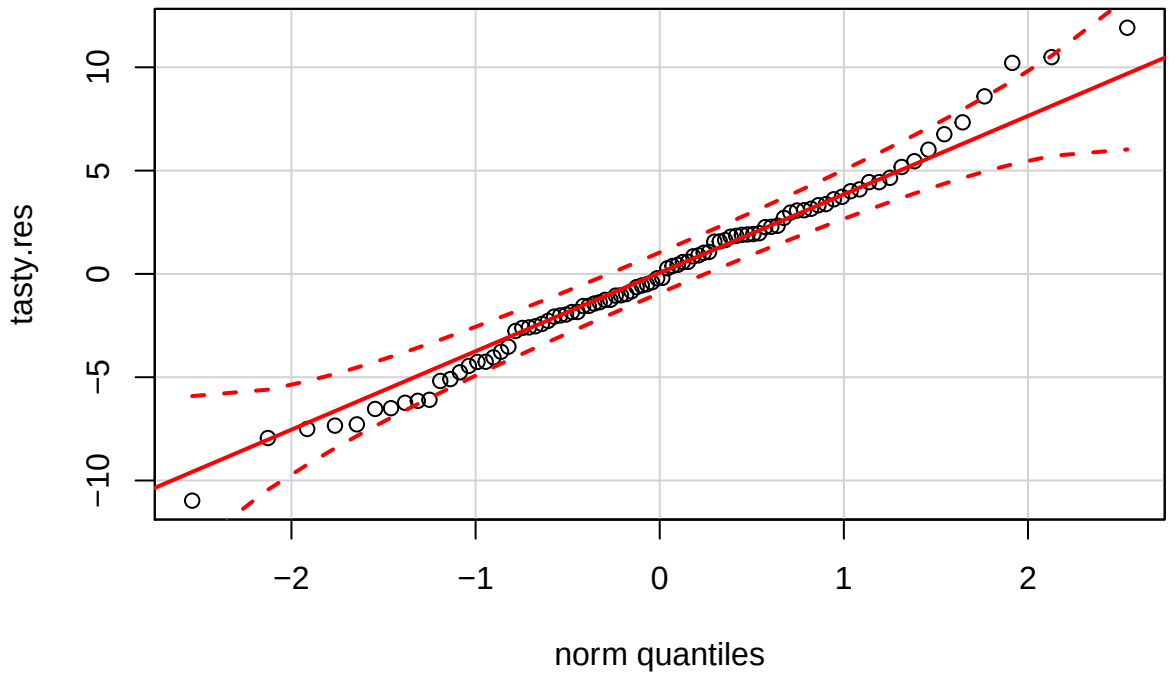
A nice Q-Q plot can be obtained by using the `qqPlot()` function in the *car* package. In addition to plotting the sample quantiles against the theoretical quantiles, it also provides **confidence bands**. If a lot of the

points on the plot fall outside of these confidence bands, it is an indication that the normality assumption may not be reasonable.

To install the *car* package, do the following: From the menu bar choose *Tools* > *Install Packages...* Type *car* under "Packages" in the window that pops up and then click on *Install*. To access functions in this package, you need to use the `library(car)` command (as is done for you below).

Run the code below to create a Q-Q plot of the residuals.

```
library(car)
qqPlot(tasty.res)
```



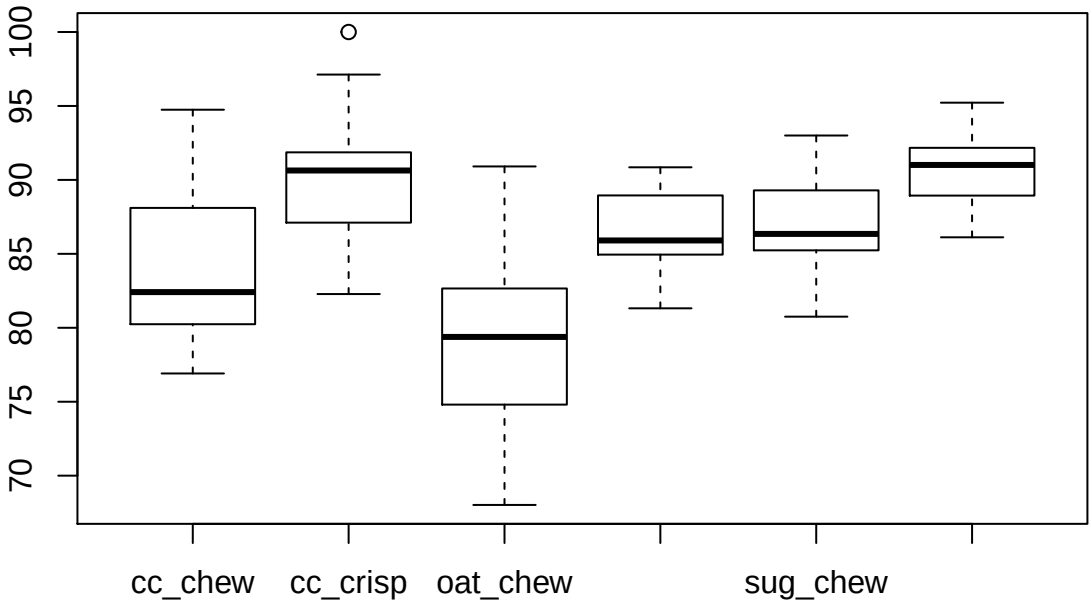
a) Does the normality assumption seem reasonable based on the Q-Q plot?

Since most of the points on this plot lie within the two dashed lines, we can assume the normality assumption is reasonable.

Boxplots and Levene's test can be used to check the equal variance assumption. For the `boxplot()` function in R, the formula for your linear model (`rating ~ cookie`) is the appropriate argument. For the Levene test, either the name of your linear model (`tasty.lm`) or the formula for the linear model (`rating ~ cookie`) can be used as the function argument. **Note: The `leveneTest()` function is also part of the *car* package in R.**

b) Create boxplots and perform Levene's test using the code chunk below.

```
boxplot(rating~cookie, data = Tastiness)
```




```
leveneTest(tasty.lm)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  5  2.2897 0.05297 .
##       84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(rating~cookie, data = Tastiness)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  5  2.2897 0.05297 .
##       84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c) Based on the p-value for the Levene test, can we reject the null hypothesis that the variance is the same for all 6 populations?

At a significance level of 0.05, we fail to reject the null hypothesis that the 6 populations have the same variance.

Problem 4

The baker from Problem 2 decided it might be best to compare mean tastiness for all pairs of cookie varieties. Tukey's HSD test limits the experimentwise type 1 error rate to be at most α when conducting all pairwise tests (and does so with less loss of power than the Bonferroni correction). The `glht()` function in the **multcomp** package will perform all pairwise tests using Tukey's method. Complete the following steps to run this test in this document.

a) Install the **multcomp** R package. This is easily done by choosing **Tools>Install Packages...**. When the window pops up, type *multcomp* into the window for the package to be installed.

b) Use the code provided below to perform the test. Remember, when you install a new package, to access functions from that package you need to first run the `library()` function (as below).

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```
tasty.tukey = glht(tasty.lm, linfct = mcp(cookie="Tukey"))
summary(tasty.tukey)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = rating ~ cookie, data = Tastiness)
##
## Linear Hypotheses:
##                            Estimate Std. Error t value Pr(>|t|)
## cc_crisp - cc_chew == 0      5.5312     1.6207   3.413  0.01245 *
## oat_chew - cc_chew == 0     -5.2587     1.6207  -3.245  0.02024 *
## oat_crisp - cc_chew == 0     2.1540     1.6207   1.329  0.76813
## sug_chew - cc_chew == 0      2.7336     1.6207   1.687  0.54436
## sug_crisp - cc_chew == 0     6.3167     1.6207   3.897  0.00263 **
## oat_chew - cc_crisp == 0   -10.7900     1.6207  -6.657  < 0.001 ***
## oat_crisp - cc_crisp == 0   -3.3772     1.6207  -2.084  0.30568
## sug_chew - cc_crisp == 0    -2.7976     1.6207  -1.726  0.51871
## sug_crisp - cc_crisp == 0    0.7855     1.6207   0.485  0.99659
## oat_crisp - oat_chew == 0    7.4128     1.6207   4.574  < 0.001 ***
## sug_chew - oat_chew == 0     7.9924     1.6207   4.931  < 0.001 ***
## sug_crisp - oat_chew == 0   11.5754     1.6207   7.142  < 0.001 ***
## sug_chew - oat_crisp == 0    0.5796     1.6207   0.358  0.99921
## sug_crisp - oat_crisp == 0   4.1627     1.6207   2.568  0.11672
## sug_crisp - sug_chew == 0    3.5831     1.6207   2.211  0.24386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

c) At an experimentwise significance level of 0.01, which pairs of means for tastiness are significantly different based on this test?

Crispy Sugar and Chewy Chocolate Chip

Chewy Oatmeal and Cripy Chocolate Chip

Chewy Oatmeal and Crispy Oatmeal

Chewy Oatmeal and Crispy Sugar

Chewy Oatmeal and Chewy Sugar