Chapter 1

Analysis of Variance

Introduction

We are now moving into a different realm of statistics. We have covered enough probability and the basic ideas of hypothesis tests and p-values to move onto the type of inference that you took this class to learn. The heart of science is comparing and evaluating which hypothesis is better supported by the data.

To evaluate a hypothesis, scientists will write a grant, hire grad students (or under-grads), collect the data, and then analyze the data using some sort of model that reflects the hypothesis under consideration. It could be as simple as "What is the relationship between iris species and petal width?" or as complex as "What is the temporal variation in growing season length in response to elevated CO_2 in desert ecosystems?"

At the heart of the question is which predictors should be included in my model of the response variable. Given twenty different predictors, I want to pare them down to just the predictors that matter. I want to make my model as simple as possible, but still retain as much explanatory power as I can.

Our attention now turns to building models of our observed data in a fashion that allows us to ask if a predictor is useful in the model or if we can remove it. Our model building procedure will be consistent:

1. Write two models, one that is perhaps overly simple and another that is a complication of the simple model.
2. Verify that the assumptions that are made in both models are satisfied.
3. Evaluate if the complex model explains significantly more of the variability in the data than the simple model.

Our goal here isn't to find "the right model" because no model is right. Instead our goal is to find a model that is *useful* and helps me to understand the science.

We will start by developing a test that helps me evaluate if a model that has a categorical predictor variable for a continuous response should have a mean value for each group or just one overall mean.

1.1 Model

The two-sample t-test provided a convenient way to compare the means from two different populations and test if they were equal. We wish to generalize this test to more than two different populations.[1]

Suppose that my data can be written as

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{where } \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma)$$

and μ_i is the mean of group i and ϵ_{ij} are the deviations from the group means. Let the first subscript denote which group the observation is from $i \in \{1, \ldots k\}$ and the second subscript is the observation number within that sample. Each group has its own mean μ_i and we might allow the number of observations in each group n_i to be of different across the populations.

Assumptions:

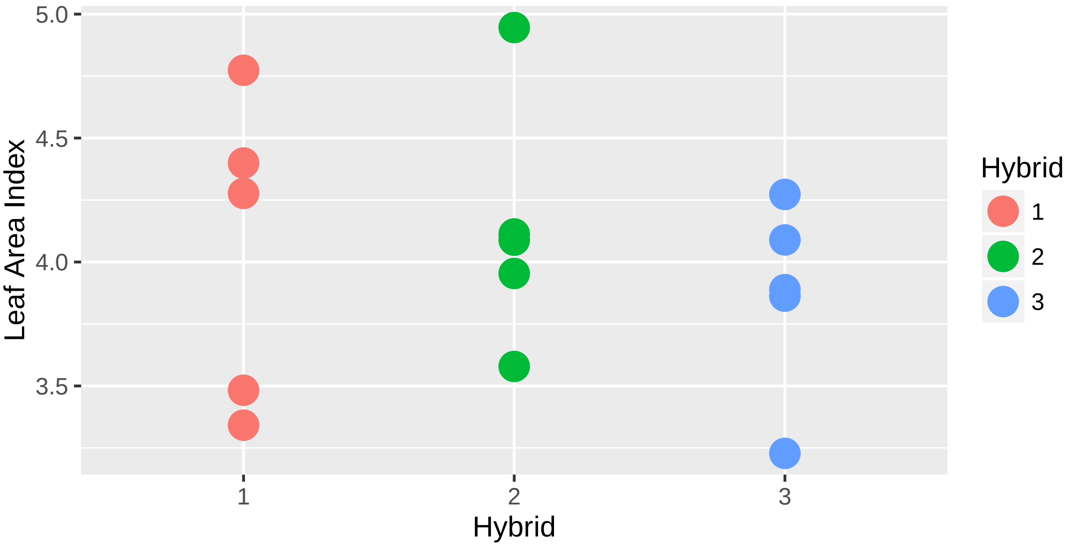
1. *The error terms come from a normal distribution*
2. *The variance of each group is the same*
3. *The observations are independent*
4. *The observations are representative of the population of interest*

In general I want to test the hypotheses

$$\begin{aligned} H_0 : & \quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a : & \quad \text{at least on mean is different than the others} \end{aligned}$$

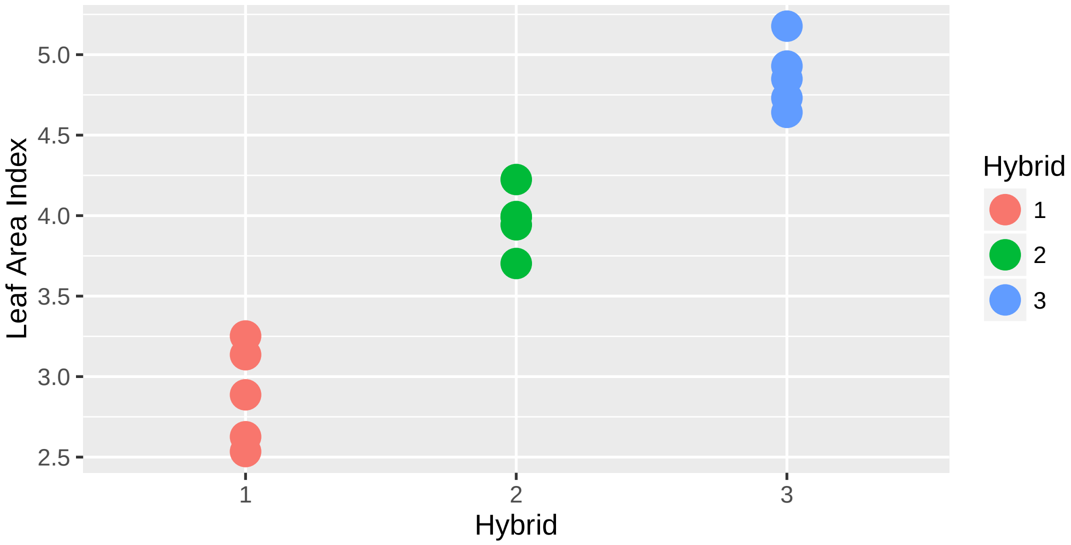
Example 1. Suppose that we have three hybrids of a particular plant and we measure the leaf area for each hybrid.

In the following graph, there does not appear to be a difference between the hybrid means:



However, in this case, it looks like there is a difference in the means of each hybrid:

[1] Later when we have more tools in our statistical tool box, it is useful to notice that ANOVA uses a categorical variable (which group) to predict a continuous response.



What is the difference between these two?

1. If the variance *between* hybrids is small compared the variance *within* a hybrid variance is huge compared, then I would fail to reject the null hypothesis of equal means (this would be the first case). In this case, the additional model complexity doesn't result in more accurate model, so Occam's Razor would lead us to prefer the simpler model where each group has the same mean.

2. If there is a large variance *between* hybrids compared to the variance *within* a hybrid then I'd conclude there is a difference (this would be the first case). In this case, I prefer the more complicated model with each group having separate means.

1.2 Theory

Notation:

1. $n=n_1+n_2+\cdots+n_k$ as the total number of observations

2. $\bar{y}_{i\cdot}=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}$ as the sample mean from the ith group

3. $\bar{y}_{\cdot\cdot}$ be the mean of all the observations.

Regardless of if the null hypothesis is true, the following is an estimate of σ^2. We could use a pooled variance estimate similar to the estimator in the pooled two-sample t-test. We will denote this first estimator as the *within-group* estimate because the sums in the numerator are all measuring the variability within a group.

$$\begin{aligned}
s_W^2 &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_k}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2}{n-k} \\
&= \frac{\sum_{j=1}^{n_1}\left(y_{1j}-\bar{y}_{1\cdot}\right)^2+\sum_{j=1}^{n_2}\left(y_{2j}-\bar{y}_{2\cdot}\right)^2+\cdots+\sum_{j=1}^{n_k}\left(y_{kj}-\bar{y}_{k\cdot}\right)^2}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)} \\
&= \frac{(n_1-1)\,s_1^2+(n_2-1)\,s_2^2+\cdots+(n_k-1)\,s_k^2}{n-k}
\end{aligned}$$

If the null hypothesis is true and $\mu_1 = \cdots = \mu_k$, then a second way that I could estimate the σ^2 is using the sample means. If H_0 is true then each sample mean has sampling distribution $\bar{Y}_{i\cdot} \sim N\left(\mu, \frac{\sigma^2}{n_i}\right)$. In the simple case where $n_1 = n_2 = \cdots = n_k$ then the sample variance of the k sample means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ has expectation σ^2/n_i and could be used to estimate σ^2. In the case of unequal sample sizes, the formula will be slightly different.

$$s_B^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2$$

Under the null hypothesis, these two estimates are both estimating σ^2 and should be similar and the ratio s_B^2/s_W^2 follows an F-distribution with numerator degrees of freedom $k-1$ and denominator degrees of freedom $n-k$ degrees of freedom. We define our test statistic as

$$f = \frac{s_B^2}{s_W^2}$$

In the case that the null hypothesis is false (non-equal means $\mu_1, \mu_2, \ldots, \mu_k$), s_B^2 should be much larger than s_W^2 and our test statistic f will be very large and so we will reject the null hypothesis if f is greater than the $1-\alpha$ quantile from the F-distribution with $k-1$ and $n_t - k$ degrees of freedom. If s_B^2 is small, then the difference between the group means and the overall means is small and we shouldn't reject the null hypothesis. So this F-test will always be a one sided test, rejecting only if f is large.

$$p - value = P\left(F_{k-1,\, n_t - k} > f\right)$$

1.2.1 Anova Table

There are several sources of variability that we are dealing with.

SSW: Sum of Squares Within - This is the variability within sample groups. It has an associated $df_W = n - k$

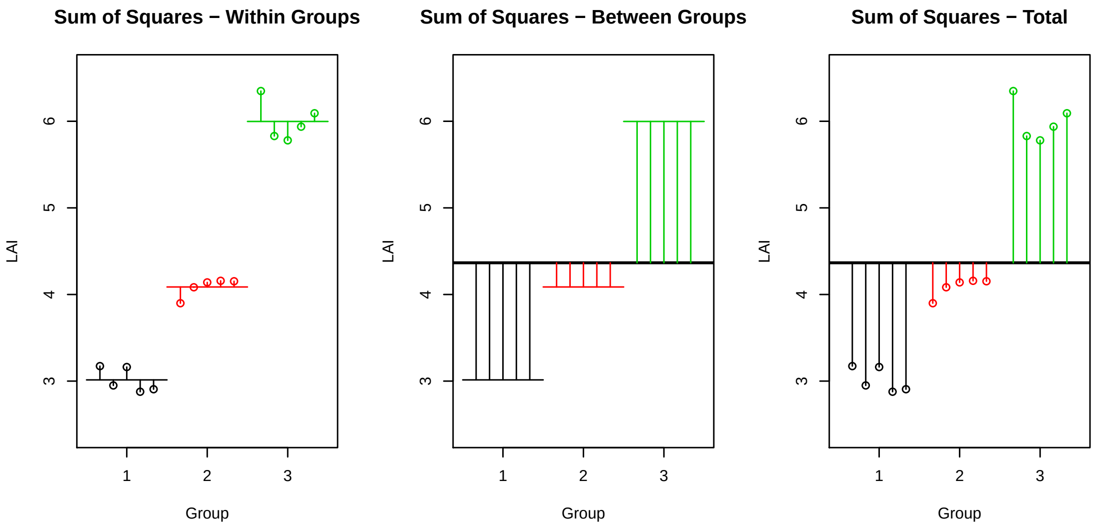
$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

SSB: Sum of Squares Between - This is the variability between sample groups. It has an associated $df_B = k - 1$

$$SSB = \sum_{i=1}^{k} n_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2$$

SST: Sum of Squares Total - This is the total variability in the data set. It has an associated $df = n - 1$ because under the null hypothesis there is only one mean μ.

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_j}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$$



An anova table is usually set up the in the following way (although the total row is sometimes removed):

Source	df	Sum of Squares	Mean Squares	F-stat	P-value
Between Samples	$k-1$	SSB	$s_B^2 = SSB/\left(k-1\right)$	$f = s_B^2/s_W^2$	$P\left(F_{k-1,n-k} > f\right)$
Within Samples	$n-k$	SSW	$s_W^2 = SSW/\left(n_t-k\right)$		
Total	$n-1$	SST			

It can be shown that

$$SST = SSB + SSW$$

and we can think about what these sums actually mean by returning to our idea about simple vs complex models.

1.2.2 ANOVA using Simple vs Complex models.[2]

The problem under consideration boils down to how complicated of a model should we fit.

Simple

The simple model is

$$Y_{ij} = \mu + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

has each observation having the same expectation μ. Thus we use the overall mean of the data $\bar{y}_{\cdot\cdot}$ as the estimate of μ and therefore our error terms are

$$e_{ij} = y_{ij} - \bar{y}_{\cdot\cdot}$$

[2] Upon the second reading of these notes, the student is likely asking why we even bothered introducing the ANOVA table using SST, SSW, SSB. The answer is that these notations are common in the ANOVA literature and that we can't justify using an F-test without variance estimates. Both interpretations are valid, but the Simple/Complex models are a better paradigm as we move forward.

The sum of squared error associated with the simple model is thus

$$\begin{aligned} SSE_{simple} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij}-\bar{y}_{\cdot\cdot}\right)^2 \\ &= SST \end{aligned}$$

Complex

The more complicated model

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \overset{iid}{\sim} N\left(0,\sigma^2\right)$$

has each observation having the expectation of it's group mean μ_i. We'll use the group means $\bar{y}_{i\cdot}$ as estimates for μ_i and thus the error terms are

$$e_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

and the sum of squared error associated with the complex model is thus

$$\begin{aligned} SSE_{complex} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij}-\bar{y}_{i\cdot}\right)^2 \\ &= SSW \end{aligned}$$

Difference

The difference between the simple and complex sums of squared error is denoted SSE_{diff} and we see

$$\begin{aligned} SSE_{diff} &= SSE_{simple} - SSE_{complex} \\ &= SST - SSW \\ &= SSB \end{aligned}$$

Note that SSE_{diff} can be interpreted as the amount of variability that is explained by the more complicated model vs the simple. If this SSE_{diff} is large, then we should use the complex model. Our only question becomes "How large is large?"

First we must account for the number of additional parameters we have added. If we added five parameters, I should expect to account for more variability that if I added one parameter, so first we will divide SSE_{diff} by the number of added parameters to get MSE_{diff} which is the amount of variability explained by each additional parameter. If that amount is large compared to the leftover from the complex model, then we should use the complex model.

These calculations are preformed in the ANOVA table, and the following table is identical to the previous ANOVA table, and we have only changed the names given to the various quantities.

Source	df	Sum of Squares	Mean Squares	F-stat	P-value
Difference	$k-1$	SSE_{diff}	$MSE_{diff}=\frac{SSE_{diff}}{k-1}$	$f=\frac{MSE_{diff}}{MSE_{complex}}$	$P\left(F_{k-1,n-k}>f\right)$
Complex	$n-k$	$SSE_{complex}$	$MSE_{complex}=\frac{SSE_{complex}}{n-k}$		
Simple	$n-1$	SSE_{simple}			

1.2.3 Parameter Estimates and Confidence Intervals

As usual, the sample mean $\bar{y}_{i\cdot}$ is a good estimator for the mean of group μ_i.

But what about σ^2? If we conclude that we should use the complex model, and since one of our assumptions is that each group has equal variance, then I should use all of the residual terms $e_{ij} = y_{ij} - \bar{y}_{i\cdot}$ in my estimation of σ. In this case we will use

$$\hat{\sigma}^2 = s_W^2 = MSE_{complex} = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

as the estimate of σ^2. Notice that this is analogous to the pooled estimate of the variance in a two-sample t-test with the assumption of equal variance.

Therefore an appropriate confidence interval for μ_i is

$$\bar{y}_{i\cdot} \pm t_{n-k}^{1-\alpha/2}\left(\frac{\hat{\sigma}}{\sqrt{n_i}}\right)$$

1.3 Anova in R

First we must define a data frame with the appropriate columns. We start with two vectors, one of which has the leaf area data and the other vector denotes the species. Our response variable must be a continuous random variable and the explanatory is a discrete variable. In R discrete variables are called `factors` and can you can change a numerical variable to be a factor using the function `factor()`.

The analysis of variance method is an example of a linear model which can be fit in a variety of ways. We can use either `lm()` or `aov()` to fit this model, and the following we will concentrate on using `aov()`. The first argument to this function is a formula that describes the relationship between the explanatory variables and the response variable. In this case it is extremely simple, that `LAI` is a function of the categorical variable `Species`.

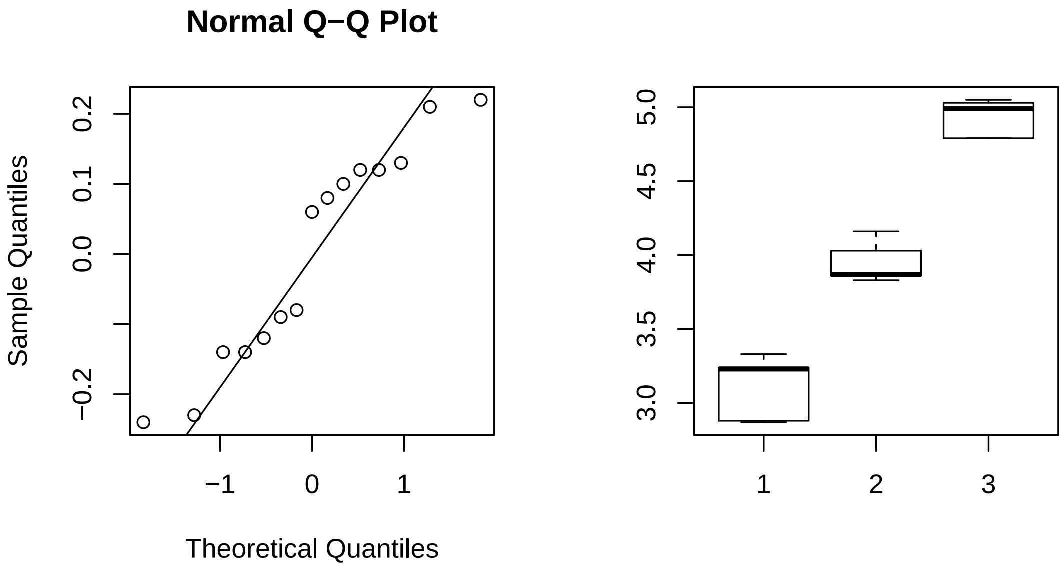
```
data <- data.frame(LAI = c(2.88, 2.87, 3.23, 3.24, 3.33,
                           3.83, 3.86, 4.03, 3.87, 4.16,
                           4.79, 5.03, 4.99, 4.79, 5.05),
                   Species = factor( rep(1:3, each=5) ) )
str(data)

## 'data.frame': 15 obs. of  2 variables:
##  $ LAI    : num  2.88 2.87 3.23 3.24 3.33 3.83 3.86 4.03 3.87 4.16 ...
##  $ Species: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 2 2 2 2 ...

model <- aov(LAI ~ Species, data=data)
```

The `aov()` command is the command that does all the calculations necessary to fit an ANOVA model. This command returns a list object that is useful for subsequent analysis and it is up to the use to know what subsequent functions to call that answer questions of interest.

In the call to `aov()` we created a formula. Formulas in R always are of the form `Y ~ X` where `Y` is the dependent variable and the `X` variables are the independent variables. In the formula we passed to `aov()`, we used a `LAI ~ Species`.

Before we examine the anova table and make any conclusion, we should double check that the anova assumptions have been satisfied. To check the normality assumption, we will look at the qqplot of the residuals $e_{ij} = y_{ij} - \bar{y}_{i\cdot}$. These residuals are easily accessed in R using the `resid` function on the object `model`. To check the variance assumption, we will examine the boxplot of the data

```
par(mfrow=c(1,2)) # side-by-side plots...
qqnorm( resid(model) )
qqline( resid(model) )
boxplot(LAI~Species, data=data)
```



The qqplot doesn't look too bad, with only two observations far from the normality line. The equal variance assumption seems acceptable as well. To get the Analysis of Variance table, we'll extract it from the `model` object using the function `anova()`.

```
model <- aov(LAI ~ Species, data=data)
anova(model)

## Analysis of Variance Table
##
## Response: LAI
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Species    2 8.2973  4.1487  147.81 3.523e-09 ***
## Residuals 12 0.3368  0.0281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that R does not give you the third line in the ANOVA table. This was a deliberate choice by the Core Development Team of R, but one that is somewhat annoying[3]. Because the third line is just the total of the first two, it isn't hard to calculate, if necessary.

The row labeled `Species` corresponds to the difference between the simple and complex models, while the `Residuals` row corresponds to the complex model. Notice that SSE_{diff} is quite large, but to decide if it is large enough to justify the use of the complex model, we must go through

[3] The package `NCStats` modifies the print command for an `aov` object to create the missing third row.

the calculations to get the p-value, which is quite small. Because the p-value is smaller than any reasonable α-level, we can reject the null hypothesis and conclude that at least one of the means is different than the others.

But which mean is different? The first thing to do is to look at the point estimates and confidence intervals for μ_i. These are

$$\hat{\mu}_i = \bar{y}_{i\cdot}$$

$$\hat{y}_{i\cdot} \pm t_{n_t-k}^{1-\alpha/2}\left(\frac{\hat{\sigma}}{\sqrt{n_i}}\right)$$

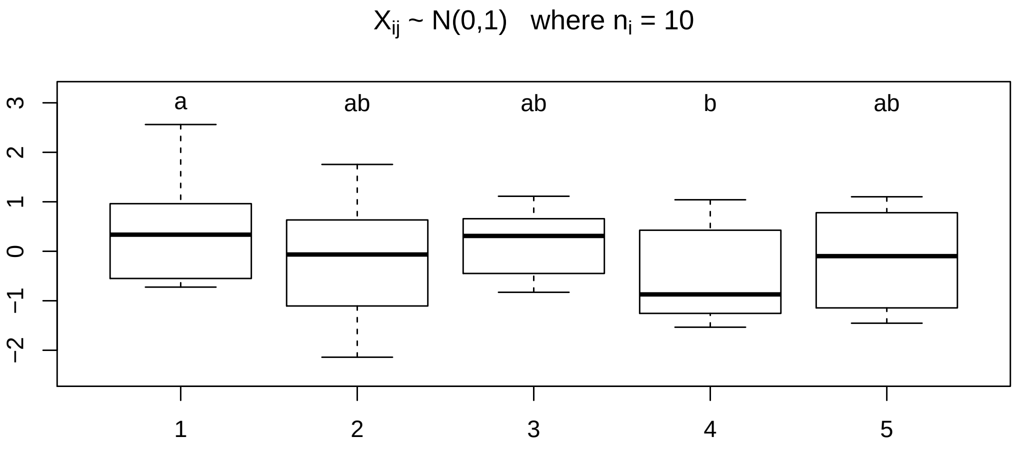
and can be found using the `coef()` and `confint()` functions.

```
# To get coefficients in the way we have represented the
# complex model (which we call the cell means model), we
# must add a -1 to the formula passed to aov()
# We'll explore this more in section 5 of this chapter.
model.2 <- aov(LAI ~ Species - 1, data=data)
coef(model.2)

## Species1 Species2 Species3
##     3.11     3.95     4.93

confint(model.2)

##             2.5 %   97.5 %
## Species1 2.946759 3.273241
## Species2 3.786759 4.113241
## Species3 4.766759 5.093241
```

Are the all the species different from each other? In practice I will want to examine each group and compare it to all others and figure out if they are different. How can we efficiently do all possible t-tests and keep the correct α level correct?

1.4 Multiple comparisons

Recall that for every statistical test there is some probability of making a type I error and we controlled that probability by setting a desired α-level. If I were to do 20 t-tests of samples with identical means, I would expect, on average, that one of them would turn up to be significantly different just by chance. If I am making a large number of tests, each with a type I error rate of α, I am practically guaranteed to make at least one type I error.



With 5 groups, there are 10 different comparisons to be made, and just by random chance, one of those comparisons might come up significant. In this sampled data, performing 10 different two sample t-tests without making any adjustments to our α-level, we find one statistically significant difference even though all of the data came from a standard normal distribution.

I want to be able to control the family-wise error rate so that the probability that I make one or more type I errors in the set of m of tests I'm considering is α. One general way to do this is called the Bonferroni method. In this method each test is performed using a significance level of α/m. (In practice I will multiple each p-value by m and compare each p-value to my desired family-wise α-level). Unfortunately for large m, this results in unacceptably high levels of type II errors. Fortunately there are other methods for addressing the multiple comparisons issue and they are built into R.

John Tukey's test of "Honestly Significant Differences" is commonly used to address the multiple comparisons issue when examining all possible pairwise contrasts. This method is available in R by the function `TukeyHSD`. This test is near optimal when each group has the same number of samples (which is often termed "a balanced design"), but becomes more conservative (fails to detect differences) as the design becomes more unbalanced. In extremely unbalanced cases, it is preferable to use a Bonferroni adjustment.

Using TukeyHSD, the adjusted p-value for the difference between groups 1 and 4 is no longer significant.

```
# TukeyHSD is very picky and will not accept Y ~ Group - 1
model <- aov(Y~Group, mydata)
TukeyHSD(model)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Y ~ Group, data = mydata)
##
## $Group
##            diff        lwr       upr     p adj
## 2-1 -0.55735682 -1.8277879 0.7130742 0.7244152
## 3-1 -0.23996214 -1.5103932 1.0304689 0.9830031
## 4-1 -0.98855350 -2.2589845 0.2818775 0.1943377
## 5-1 -0.62440394 -1.8948350 0.6460271 0.6330050
## 3-2  0.31739468 -0.9530364 1.5878257 0.9531756
## 4-2 -0.43119668 -1.7016277 0.8392344 0.8695429
## 5-2 -0.06704712 -1.3374782 1.2033839 0.9998817
## 4-3 -0.74859136 -2.0190224 0.5218397 0.4596641
## 5-3 -0.38444180 -1.6548728 0.8859892 0.9099064
## 5-4  0.36414956 -0.9062815 1.6345806 0.9248234
```

Likewise if we are testing the ANOVA assumption of equal variance, we cannot rely on doing all pairwise F-tests and we must use a method that controls the overall error rate. The multiple comparisons version of `var.test()` is Bartlett's test which is called similarly to `aov()`.

```
bartlett.test(Y~Group, mydata)

##
##  Bartlett test of homogeneity of variances
##
## data:  Y by Group
## Bartlett's K-squared = 3.1397, df = 4, p-value = 0.5347
```

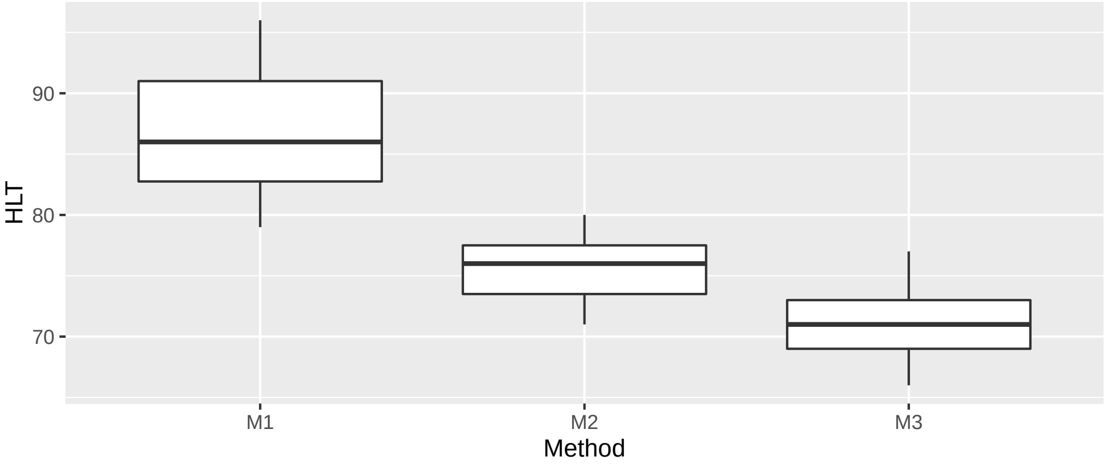
Example 2. (Example 8.2 from the Ott and Longnecker) A clinical psychologist wished to compare three methods for reducing hostility levels in university students, and used a certain test (HLT) to measure the degree of hostility. A high score on the test indicated great hostility. The psychologist used 24 students who obtained high and nearly equal scores in the experiment. Eight subjects were selected at random from among the 24 problem cases and were treated with method 1, seven of the remaining 16 students were selected at random and treated with method 2 while the remaining nine students were treated with method 3. All treatments were continued for a one-semester period. Each student was given the HLT test at the end of the semester, with the results show in the following table. Use these dat to perform an analysis of variance to determine whether there are differences among the mean scores for the three methods using a significance level of $\alpha = 0.05$.

Method	Test Scores								
1	96	79	91	85	83	91	82	87	
2	77	76	74	73	78	71	80		
3	66	73	69	66	77	73	71	70	74

```
# define the data
Hostility <- data.frame(
  HLT = c(96,79,91,85,83,91,82,87,
          77,76,74,73,78,71,80,
          66,73,69,66,77,73,71,70,74),
  Method = c( rep('M1',8), rep('M2',7), rep('M3',9) ) )
```

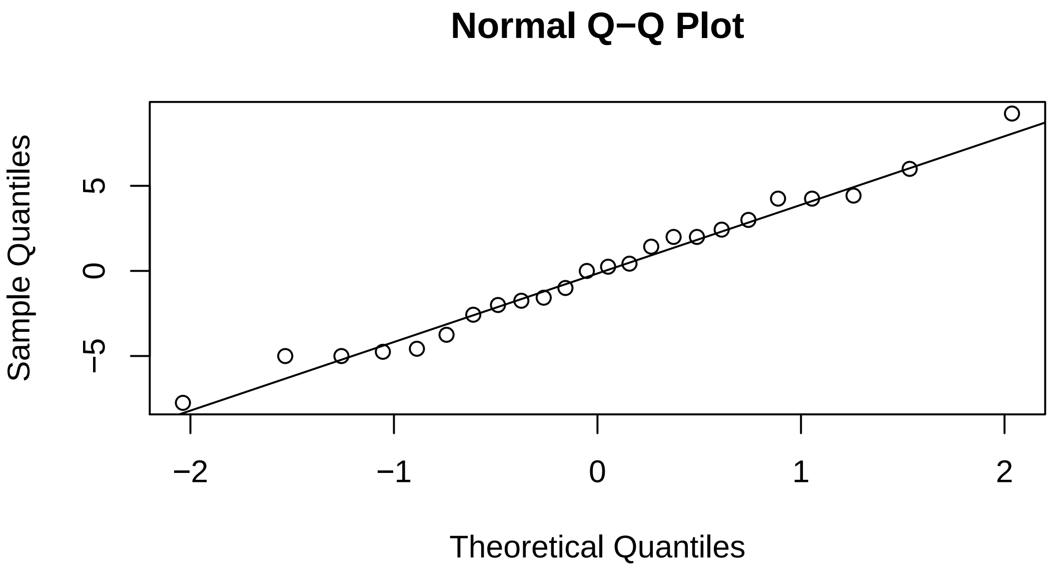
The first thing we will do (as we should do in all data analyses) is to graph our data.

```
library(ggplot2)
ggplot(Hostility, aes(x=Method, y=HLT)) +
  geom_boxplot()
```



These box plots make it clear that there is a difference between the three groups (at least group `M1` is different from `M2` or `M3`). An ANOVA model assumes equal variance between groups and that the residuals are normally distributed. Based on the box plot, the equal variance assumption might be suspect (although with only ≈ 8 observations per group, it might not be bad). We'll examine a QQ-plot of the residuals to consider the normality.

```
# Do the model assumptions hold?
model <- aov( HLT ~ Method, data=Hostility )
qqnorm( resid(model) )
qqline( resid(model) )
```



To examine the Normality of the residuals, we'll use a Shapiro-Wilk's test and we'll also use Bartlett's test for homogeneity of variances.

```
# Test for Normality
shapiro.test(resid(model))

##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.98358, p-value = 0.9516

# Test for equal variances between groups
bartlett.test(HLT~Method, data=Hostility)

##
##  Bartlett test of homogeneity of variances
##
## data:  HLT by Method
## Bartlett's K-squared = 2.4594, df = 2, p-value = 0.2924
```

The results of the Shapiro-Wilks test agree with the QQ-plot, and Bartlett's test fails to detect differences in the variances between the two groups. This is not to say that there might not be a difference, only that we do not detect one.

```
model <- aov( HLT ~ Method, data=Hostility )
anova(model)

## Analysis of Variance Table
##
## Response: HLT
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Method     2 1090.62  545.31  29.574 7.806e-07 ***
## Residuals 21  387.21   18.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value in the ANOVA table is smaller than $\alpha = 0.05$, we can reject the null hypothesis of equal means and conclude that at least one of the means is different from the others. Our estimate of σ^2 is 18.44 so the estimate of $\sigma = \sqrt{18.44} = 4.294$.

To find out which means are different we first look at the group means and confidence intervals.

```
# To get the group means from aov, we must
# use the -1 in the formula command
model.2 <- aov( HLT ~ Method - 1, data=Hostility )
coef(model.2)

## MethodM1 MethodM2 MethodM3
## 86.75000 75.57143 71.00000

confint(model.2)

##             2.5 %   97.5 %
## MethodM1 83.59279 89.90721
## MethodM2 72.19623 78.94663
## MethodM3 68.02335 73.97665
```

To control for the multiple comparisons issue we again look at all possible group comparisons using the `TukeyHSD` function.

```
# Remember TukeyHSD is picky and doesn't like the -1...
TukeyHSD(model)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = HLT ~ Method, data = Hostility)
##
## $Method
##             diff       lwr         upr     p adj
## M2-M1 -11.178571 -16.78023  -5.5769151 0.0001590
## M3-M1 -15.750000 -21.00924 -10.4907592 0.0000006
## M3-M2  -4.571429 -10.02592   0.8830666 0.1113951
```

If we feel uncomfortable with the equal variance assumption, we can do each pairwise t-test using non-pooled variance and then correct for the multiple comparisons using Bonferroni's p-value correction. If we have $k = 3$ groups, the we have $k(k-1)/2 = 3$ different comparisons, so I will calculate each p-value and multiply by 3.

```
pairwise.t.test(Hostility$HLT, Hostility$Method,
                pool.sd=FALSE, p.adjust.method='none')

##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  Hostility$HLT and Hostility$Method
##
##    M1      M2
## M2 0.0005  -
## M3 2.2e-05 0.0175
##
## P value adjustment method: none

pairwise.t.test(Hostility$HLT, Hostility$Method,
                pool.sd=FALSE, p.adjust.method='bonferroni')

##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  Hostility$HLT and Hostility$Method
##
##    M1      M2
## M2 0.0015  -
## M3 6.7e-05 0.0525
##
## P value adjustment method: bonferroni
```

Using the Bonferroni adjusted p-values, we continue to detect a statistically significant difference between Method 1 and both Methods 2 & 3, but do not detect a difference between Method 2 and Method 3.

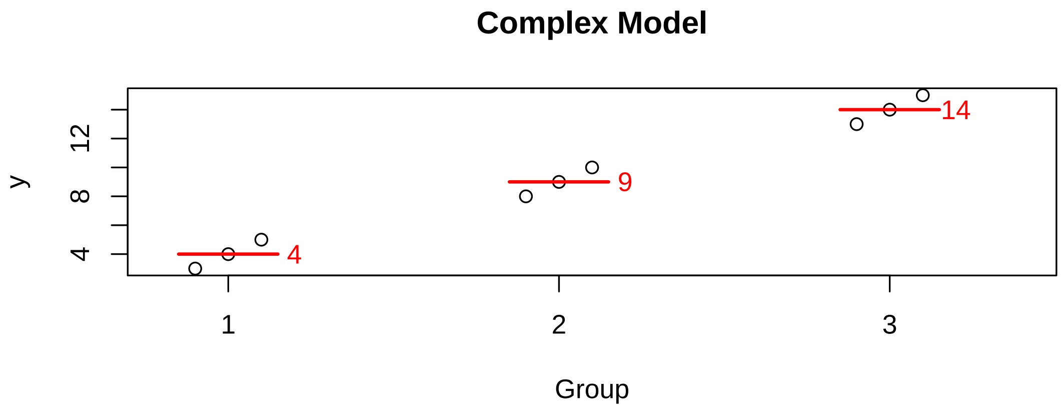
1.5 Different Model Representations

1.5.1 Theory

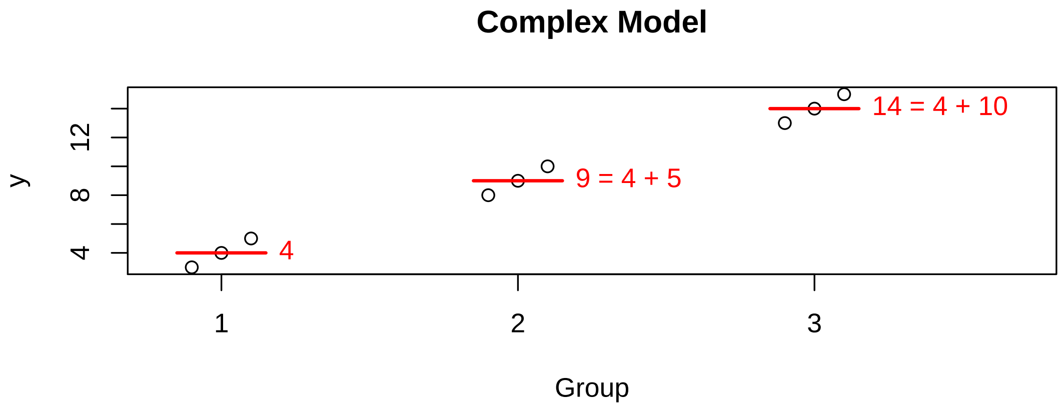
We started with what I will call the “cell means model”

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{where} \quad \epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

so that the $E\left(Y_{ij}\right) = \mu_i$ where I interpret μ_i as the mean of each population. Given some data, we the following graph where the red lines and numbers denote the observed mean of the data in each group :



But I am often interested in the difference between one group and another. For example, suppose this data comes from an experiment and group 1 is the control group. Then perhaps what I'm really interested is not that group 2 has a mean of 9, but rather that it is 5 units larger than the control. In this case perhaps what we care about is the differences. I could re-write the group means in terms of these differences from group 1. So looking at the model this way, the values that define the group means are the mean of group 1 (here it is 4), and the offsets from group 1 to group 2 (which is 5), and the offset from group 1 to group 3 (which is 10).



I could write this interpretation of the model as the "offset" model which is

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where μ is the mean of group 1 and τ_i is each population's offset from group 1. Since group 1 can't be offset from itself, this forces $\tau_1 = 0$.

Notice that this representation of the complex model has 4 parameters (aside from σ), but it has an additional constraint so we still only have 3 parameters that can vary (just as the cell means model has 3 means).

The cell means model and the offset model really are the same model, just looked at slightly differently. They have the same number of parameters, and produce the same predicted values for $\hat{y}_{ij}$ and therefore have the same sum of squares, etc. The only difference is that one is might be more convenient depending on the question the investigator is asking.

Another way to write the cell means model is as

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

but with the constraint that $\mu = 0$. It doesn't matter which constraint you use so long as you know which is being used because the interpretation of the values changes (group mean versus an offset from the reference group).

1.5.2 Model Representations in R

To obtain the different representations within R, we will vary the formula to include or exclude the intercept term μ. By default, R assumes you want the intercept term (offset representation) and you must use the `-1` term in the formula for the cell means representation.

```
fake.data <- data.frame(   y =        c( 3,4,5,  8,9,10, 13,14,15),
                         grp = factor(c( 1,1,1,  2,2,2,    3,3,3 )) )
# Offset representation
#   Unless you have a -1, R implicitly
#   adds a "+1" to the formula, so
#   so the following statements are equivalent
#c.model.1 <- aov(y ~ grp  , data=fake.data)
 c.model.1 <- aov(y ~ grp+1, data=fake.data)
coef(c.model.1)

## (Intercept)        grp2        grp3
##           4           5          10
```

In the above case, we see R is giving the mean of group 1 and then the two offsets.

To force R to use the cell means model, we force R to use the constraint that $\mu = 0$ by including a `-1` in the model formula.

```
c.model.1 <- aov(y ~ grp -1, data=fake.data)
coef(c.model.1)

## grp1 grp2 grp3
##    4    9   14
```

Returning the hostility example, recall we used the cell means model and we can extract parameter coefficient estimates using the `coef` function and ask for the appropriate confidence intervals using `confint()`.

```
model <- aov(HLT ~ Method - 1, data=Hostility)
coef(model)

## MethodM1 MethodM2 MethodM3
## 86.75000 75.57143 71.00000

confint(model)

##             2.5 %   97.5 %
## MethodM1 83.59279 89.90721
## MethodM2 72.19623 78.94663
## MethodM3 68.02335 73.97665
```

We can use the intercept model by removing `-1` term from the formula.

```
model <- aov(HLT ~ Method, data=Hostility)
coef(model)

## (Intercept)    MethodM2    MethodM3
##    86.75000   -11.17857   -15.75000

confint(model)

##                 2.5 %     97.5 %
## (Intercept)  83.59279  89.907212
## MethodM2    -15.80026  -6.556886
## MethodM3    -20.08917 -11.410827
```

The intercept term in the offset representation corresponds to `Method1` and the coefficients and confidence intervals are the same as in the cell means model. However in the offset model, `Method2` is the *difference* between `Method1` and `Method2`. Notice the coefficient is negative, thus telling us that `Method2` has a smaller mean value than the reference group `Method1`. Likewise `Method3` has a negative coefficient indicating that the `Method3` group is lower than the reference group.

Similarly the confidence intervals for `Method2` and `Method3` are now confidence intervals for the *difference* between these methods and the reference group `Method1`.

Why would we ever want the offset model vs the cell means model? Often we are interested in testing multiple treatments against a control group and we only care about the change from the control. In that case, setting the control group to be the reference makes sense.

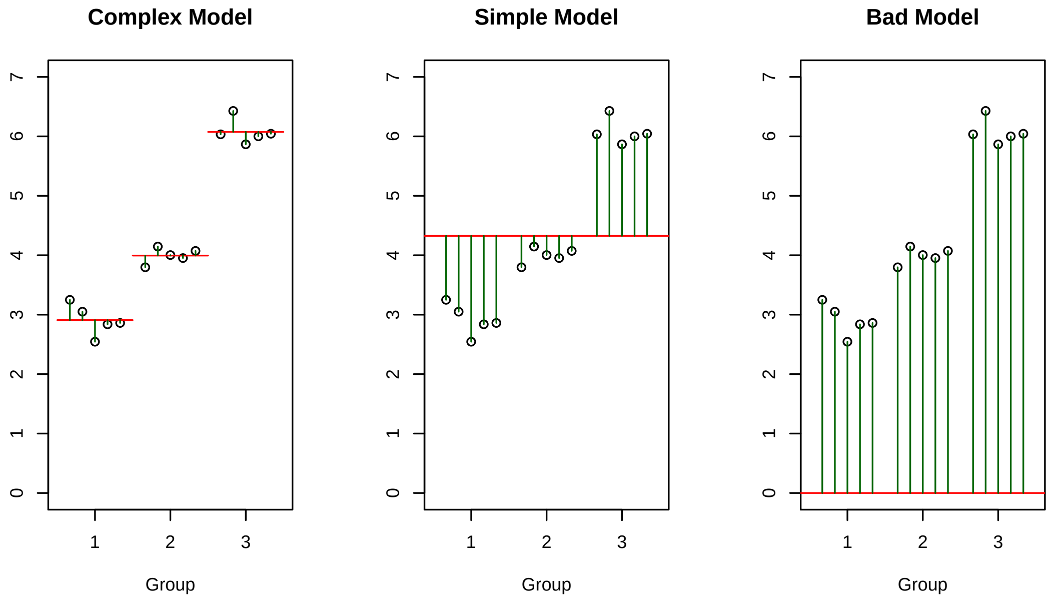
Neither representation is more powerful because on a very deep mathematical level, they are exactly the same model. Superficially though, one representation might be more convenient than the other in a given situation.

1.5.3 Implications on the ANOVA table

We have been talking about the complex and simple models for our data but there is one more possible model, albeit not a very good one. I will refer to this as the **bad model** because it is almost always a poor fitting model.

$$Y_{ij} = \epsilon_{ij}$$

where $\epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$.



Notice that the complex model has three parameters that define "signal" part of the model (i.e. the three group means). The simple has one parameter that defines the "signal" (the overall mean). The bad model has *no* parameters that define the model (i.e. the red line is always at zero).

These three models can be denoted in R by:

- Complex:
  - offset representation: `Y ~ group` which R will recognize as `Y ~ group + 1`
  - cell means representation: `Y ~ group - 1`
- Simple: `Y ~ 1`
- Bad: `Y ~ -1`

In the analysis of variance table calculated by `anova()`, R has to decide which simple model to compare the complex model to. If you used the offset representation, then when `group` is removed from the model, we are left with the model `Y ~ 1`, which is the simple model. If we wrote the complex model using the cell means representation, then when `group` is removed, we are left with the model `Y ~ -1` which is the bad model.

When we produce the ANOVA table compare the complex to the bad model, the difference in number of parameters between the models will be 3 (because I have to add three parameters to go from a signal line of 0, to three estimated group means). The ANOVA table comparing simple model to the complex will have a difference in number of parameters of 2 (because the simple mean has 1 estimated value compared to 3 estimated values).

Example. Hostility Scores

We return to the hostility scores example and we will create the two different model representations in R and see how the ANOVA table produced by R differs between the two.

```
offset.representation <- aov(HLT ~ Method, data=Hostility)
cell.representation   <- aov(HLT ~ Method -1, data= Hostility)
#
#
# This is the ANOVA table we want, comparing Complex to Simple
# Notice the df of the difference between the models is 3-1 = 2
anova(offset.representation)

## Analysis of Variance Table
##
## Response: HLT
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Method     2 1090.62  545.31  29.574 7.806e-07 ***
## Residuals 21  387.21   18.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#
#
# This is the ANOVA table comparing the Complex to the BAD model
# Noice the df of the difference between the models is 3-0 = 3
anova(cell.representation)

## Analysis of Variance Table
##
## Response: HLT
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Method     3 145551   48517  2631.2 < 2.2e-16 ***
## Residuals 21    387      18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the bad model is *extremely* bad in this case, the F-statistic for comparing the complex to the bad model is extremely large (F=2631). The complex model is also superior to the simple model, but not by as emphatically (F=29).

One way to be certain which models you are comparing is to explicitly choose the two models.

```
simple <- aov(HLT ~ 1, data=Hostility)

# create the ANOVA table comparing the complex model (using the
# cell means representation) to the simple model.
# The output shown in the following contains all the
# necessary information, but is arranged slightly differently.
anova(simple, cell.representation)

## Analysis of Variance Table
##
## Model 1: HLT ~ 1
## Model 2: HLT ~ Method - 1
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     23 1477.83
## 2     21  387.21  2    1090.6 29.574 7.806e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.6 Exercises

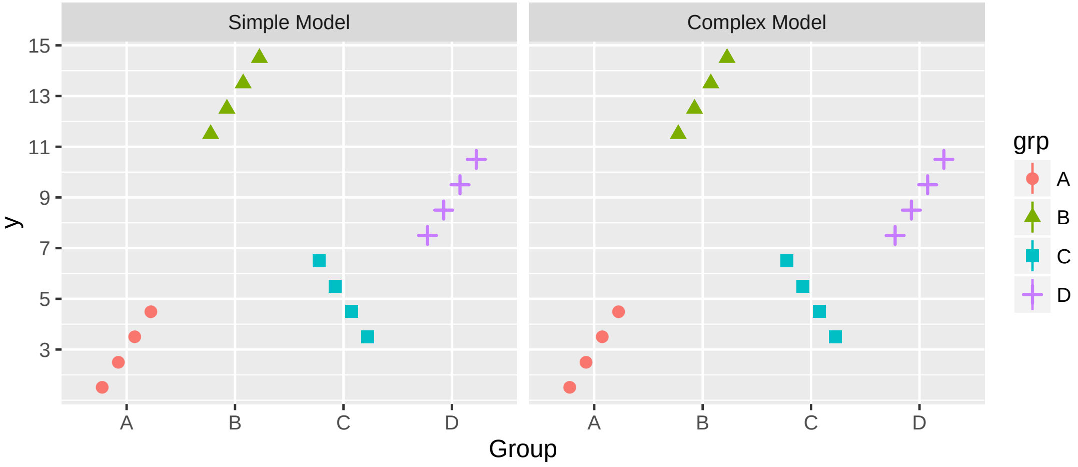
1. For this exercise, we will compare the Sums of Squared Error for the simple

$$y_{ij} = \mu + \epsilon_{ij}$$

and complex

$$y_{ij} = \mu_i + \epsilon_{ij}$$

model and clearly, in the data presented below, the complex model fits the data better. The group means $\bar{y}_{i\cdot}$ are 3, 13, 5, and 9, while the overall mean is $\bar{y}_{\cdot\cdot} = 7.5$.

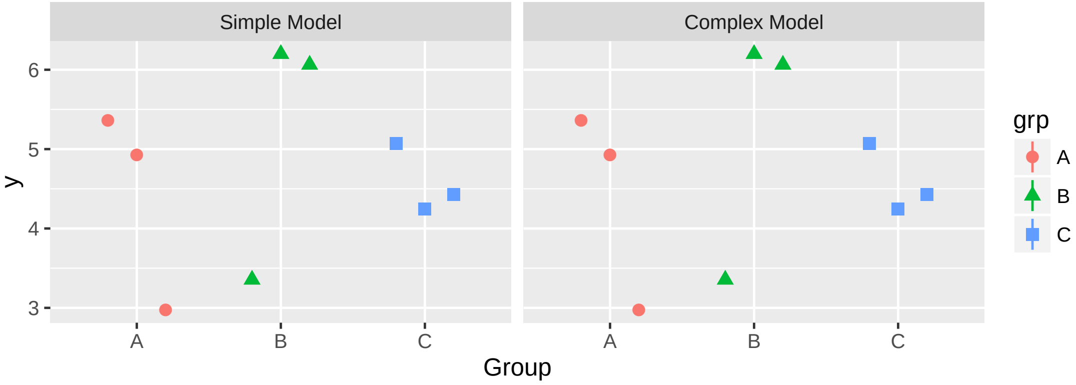


(a) For the simple model graph, draw a horizontal line at the height of the overall mean, representing predicted value of a new observation. Next, draw the the corresponding residuals $y_{ij} - \bar{y}_{\cdot\cdot}$ as vertical lines from the data points to the overall mean. Similarly draw horizontal lines for the group means in the complex model and represent the residuals for the complex model $y_{ij} - \bar{y}_{i\cdot}$ as vertical lines from the data points to the group means. In this case, does it appear that the average residual is significantly larger in the simple model than the complex?

(b) To show that the complex is a significantly better model, fill in the empty boxes in the ANOVA table.

Source	Sum of Squares	df	Mean Squares	F-stat	P-value
Difference					
Complex	20	12			
Simple	256	15			

2. We will essentially repeat the previous exercise, except this time, the simple model will be preferred.



For this data, the following means can be calculated:

```
library(dplyr)

##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:plyr':
##
##    arrange, count, desc, failwith, id, mutate, rename, summarise,
##    summarize
## The following objects are masked from 'package:stats':
##
##    filter, lag
## The following objects are masked from 'package:base':
##
##    intersect, setdiff, setequal, union

str(p2.data)

## 'data.frame': 9 obs. of  2 variables:
##  $ y  : num  5.36 4.93 2.97 3.36 6.2 ...
##  $ grp: Factor w/ 3 levels "A","B","C": 1 1 1 2 2 2 3 3 3

# calculate the group means
p2.data %>% group_by(grp) %>% summarise(xbar.i = mean(y))

## Source: local data frame [3 x 2]
##
##      grp   xbar.i
##   (fctr)    (dbl)
## 1      A 4.419685
## 2      B 5.205504
## 3      C 4.579385

# calculate the overall mean
p2.data %>% summarise(xbar = mean(y))

##       xbar
## 1 4.734858
```

(a) For the simple model graph, draw the corresponding residuals $y_{ij} - \bar{y}_{\cdot\cdot}$ as vertical lines from the data point to the overall mean. Similarly draw the residuals for the complex model $y_{ij} - \bar{y}_{i\cdot}$ as vertical lines from the data points to the group means. (Visually estimate the group and overall means). In this case, does it appear that the average residual is significantly larger in the simple model than the complex?

(b) To show that the complex not a significantly better model, fill in the empty boxes in the ANOVA table.

Source	Sum of Squares	df	Mean Squares	F-stat	P-value
Difference					
Complex	8.75				
Simple	9.79	8			

3. The following data were collected and we wish to perform an analysis of variance to determine if the group means are statistically different.

Group 1	Group 2	Group 3
$4, 6, 6, 8$	$8, 8, 6, 6$	$12, 13, 15, 16$

(a) The complex model assumes different means for each group. That is

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Calculate $SSE_{complex}$ via the following:

i. Find the estimate of μ_i. That is, calculate $\hat{\mu}_i = \bar{y}_{i\cdot}$ which is the mean of each group. Therefore the predicted value for a new observation in group i would be $\hat{y}_{ij} = \hat{\mu}_i = \bar{y}_{i\cdot}$ and you can now calculate $SSE_{complex}$.

ii. Calculate

$$SSE_{complex} = \sum_{i=1}^{3}\sum_{j=1}^{4}\left(y_{ij} - \hat{y}_{ij}\right)^2 = \sum_{i=1}^{3}\sum_{j=1}^{4}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$$

(b) The simple model assumes the same mean for each group. That is

$$Y_{ij} = \mu + \epsilon_{ij}$$

Calculate SSE_{simple} via the following:

i. Find the estimate of μ. That is, calculate $\hat{\mu} = \bar{y}_{\cdot\cdot}$ which is the overall mean of all the data. Therefore the predicted value for a new observation in any group would be $\hat{y}_{ij} = \hat{\mu} = \bar{y}_{\cdot\cdot}$ and we can calculate SSE_{simple}

ii. Calculate

$$SSE_{simple} = \sum_{i=1}^{3}\sum_{j=1}^{4}\left(y_{ij} - \hat{y}_{ij}\right)^2 = \sum_{i=1}^{3}\sum_{j=1}^{4}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$$

(c) Recreate the ANOVA table using R by typing in the data set and fitting the appropriate model using the `aov()` command. Obtain the ANOVA table from this model by using the `anova()` command.

4. Suppose that for a project I did four separate t-tests and the resulting p-values were

p_1	p_2	p_3	p_4
0.03	0.14	0.01	0.001

   If I wanted to control my overall type I error rate at an $\alpha = 0.05$ and used the Bonferroni multiple comparisons procedure, which tests would be statistically significant?

5. We will examine the amount of waste produced at five different plants that manufacture Levi Jeans. The `Waste` amount is the amount of cloth wasted in cutting out designs compared to a computer program, so negative values for `Waste` indicate that the human engineer did a better job planning the cuts than the computer algorithm. The data are available in the file "Levi.csv" available on Bblearn. There are two columns, `Plant` and `Waste`.

   (a) Read the data into R. If you need more help on this, see the chapter on importing data in the R manual: `https://raw.github.com/dereksonderegger/STA_570L_Book/master/Introduction_to_R.pdf`.

      i. Download the file `Levi.csv` from GitHub `https://raw.github.com/dereksonderegger/STA_570_Book/master/data/Levi.csv` to some location on your computer (wherever you store RMarkdown files for your homework assignments). Open this file using a simple editor (Notepad on Windows or TextEdit on a Mac) and notice that the first line is a set of column headers and subsequent lines have the data. Also notice that the columns are separated by commas. This file type is known as a "comma separated values" file and MS Excel can open it just fine and Excel can save files in this file type as well.

      ii. Make sure the working directory in R is set to the same directory. In RStudio, this is under the menu item: Tools -> Set Working Directory.

      iii. Read in the file using the following command:

```
# First term is the file name,
# second is that the first row is the column headers,
# last is that a comma separates the columns from one another
Levi <- read.table('Levi.csv', header=TRUE, sep=',')
```

      iv. Examine the data frame using the `str(Levi)` command. Is the `Plant` column already a factor, or do you need to convert it to a factor?

   (b) Make a boxplot of the data. Do any assumptions necessary for ANOVA appear to be violated?

   (c) Test the equal variance assumption using Bartlett's test.

   (d) Fit an ANOVA model to these data and test if the residuals have a normal distribution using the Shapiro-Wilks test.

6. The dataset `iris` is available on R and can be loaded by the entering the command `data('iris')` at your R prompt. This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for $n_i = 50$ flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*. We will be examining the relationship between sepal width and the species of these irises.

   Denote the mean value of all *setosa* flowers as μ_{setosa} and similar notation for the other species.

   (a) Make a boxplot of the data. Do any assumptions necessary for ANOVA appear to be violated?

(b) Test the equal variance assumption of ANOVA using Bartlett's test.

(c) Do the ANOVA test using the command

```
model <- aov( Sepal.Width ~ Species, data=iris)
```

and test the normality of the residual terms by making a QQplot and doing the Shapiro-Wilk's test.

(d) Examine the ANOVA table. What is the p-value for testing the hypotheses

$$\begin{aligned} H_0: &\quad \mu_{setosa} = \mu_{virginica} = \mu_{versicolor} \\ H_{a:} &\quad \text{at least on mean is different} \end{aligned}$$

(e) Now that we know there is a statistically significant difference amongst the means (and with *setosa* having a mean `Sepal.Width` about 30% larger than the other two, I think aesthetically that is a big difference), we can go searching for it. Use Tukey's "Honestly Significant Differences" method to test all the pairwise comparisons between means. In particular, what is the p-value for testing

$$\begin{aligned} H_0: &\quad \mu_{setosa} = \mu_{virginica} \\ H_a: &\quad \mu_{setosa} \neq \mu_{virginica} \end{aligned}$$

(f) Refit the model using the "`-1`" term

```
model.2 <- aov( Sepal.Width ~ -1 + Species, data=iris)
```

(g) Using this second model and the function `coef()`, what is the estimated value of μ_{setosa}? What is the estimated value of $\mu_{virginica}$?

(h) What is the estimated value of σ^2?

(i) By hand, calculate the appropriate 95% confidence interval for μ_{setosa}.

(j) Using the R function `confint()`, confirm your calculation in part (h).