Recap: One-Way Anova

We will generate artificial data fitting the model:

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

With:

- $\mu = 2$
- $\alpha_1 = 4$
- $\alpha_2 = 1$
- $\alpha_3 = 2$
- $\sigma_e = 10$

One-Way ANOVA

Enter the following R statements:

```
mu<-1; alpha1<-1; alpha2<-4; alpha3<-6
x1<-c(rep(1,1000),rep(0,1000),rep(0,1000))
x2<-c(rep(0,1000),rep(1,1000),rep(0,1000))
x3<-c(rep(0,1000),rep(0,1000),rep(1,1000))
e<-rnorm(3000,0,5)
group<-gl(3,1000,3000,labels=c("G1","G2","G3"))
y<-mu+alpha1*x1+alpha2*x2+alpha3*x3+e
art1<-data.frame(y,x1,x2,x3,group)
str(art1)
```

One-Way ANOVA

We can produce a boxplot of the data with the following statement:

```
boxplot(y ~ group)
```

One-Way ANOVA

We can produce a boxplot of the data with the following statement:

```
boxplot(y ~ group)
```

Now run the `aov` procedure and print the summary of its output:

```
lm0<-aov( y ~ group) ; summary(lm0)
```

One-Way ANOVA

We can produce a boxplot of the data with the following statement:

```
boxplot(y ~ group)
```

Now run the `aov` procedure and print the summary of its output:

```
lm0<-aov( y ~ group) ; summary(lm0)
```

On the line beginning with `group`, the `F value` and `Pr(>F)` indicate whether there are any significant differences between groups.

One-Way ANOVA

We can produce a boxplot of the data with the following statement:

```
boxplot(y ~ group)
```

Now run the `aov` procedure and print the summary of its output:

```
lm0<-aov( y ~ group) ; summary(lm0)
```

On the line beginning with `group`, the `F value` and `Pr(>F)` indicate whether there are any significant differences between groups.

If `Pr(>F)` is less than the desired α level of the test (usually 0.05), we reject the mull hypothesis that the group means are all equal.

One-Way ANOVA

The means of the variables y, $x1$, $x2$, and $x3$ by group can be obtained by the following statements:

```
aggregate(art1, by=list(group), FUN=mean)
```

One-Way ANOVA

The means of the variables y, $x1$, $x2$, and $x3$ by group can be obtained by the following statements:

```
aggregate(art1, by=list(group), FUN=mean)
```

From the way we generated the data, these means represent sample estimates of the following parameter values:

- $E(Y)$ for group 1: $\mu + \alpha_1 = 1 + 1 = 2$
- $E(Y)$ for group 2: $\mu + \alpha_2 = 1 + 4 = 5$
- $E(Y)$ for group 3: $\mu + \alpha_3 = 1 + 6 = 7$

One-Way ANOVA

Now run the `lm` procedure and print the summary of its output:

```
lm0<-lm( y ~ group) ; summary(lm0)
```

One-Way ANOVA

Now run the `lm` procedure and print the summary of its output:

```
lm0<-lm( y ~ group) ; summary(lm0)
```

The numbers in the `Estimate` column (not produced by the `aov` function) represents the following in terms of the parameters:

Row	Estimate	Expected Value
(Intercept)	$\mu + \alpha_1$	$1 + 1 = 2$
groupG2	$\alpha_2 - \alpha_1$	$5 - 2 = 3$
groupG3	$\alpha_3 - \alpha_1$	$7 - 2 = 5$

Reading the EPA data into R

Go to the course web page, then the *Notes and Handouts* section.

Reading the EPA data into R

Go to the course web page, then the *Notes and Handouts* section.

Right click on the *2009 EPA Mileage Data* link and select *copy link location*

Reading the EPA data into R

Go to the course web page, then the *Notes and Handouts* section.

Right click on the *2009 EPA Mileage Data* link and select *copy link location*

This should copy the URL for the EPA .csv data file, which is:

http://www.sandgquinn.org/stonehill/MA225/notes/09tstcar.csv

Reading the EPA data into R

Go to the course web page, then the *Notes and Handouts* section.

Right click on the *2009 EPA Mileage Data* link and select *copy link location*

This should copy the URL for the EPA .csv data file, which is:

http://www.sandgquinn.org/stonehill/MA225/notes/09tstcar.csv

Carefully type the following command in R, but don't hit enter:

```
epa<-read.table("",sep=",",fill=TRUE,header=TRUE)
```

One-Way ANOVA: Cylinders

We will use a one-way ANOVA to compare city mileage of cars with 4, 6, and 8 cylinders.

One-Way ANOVA: Cylinders

We will use a one-way ANOVA to compare city mileage of cars with 4, 6, and 8 cylinders.

First we will create a new dataframe called `epa468` containing only city mileage values for vehicles with 4, 6, or 8 cylinders:

```
epa468<- subset(epa, C.H=="C" & (vpc==4 |
vpc==6 | vpc==8))
```

One-Way ANOVA: Cylinders

We will use a one-way ANOVA to compare city mileage of cars with 4, 6, and 8 cylinders.

First we will create a new dataframe called `epa468` containing only city mileage values for vehicles with 4, 6, or 8 cylinders:

```
epa468<- subset(epa, C.H=="C" & (vpc==4 |
vpc==6 | vpc==8))
```

Next we select only records for cars, and keep only mpg and vpc:

```
epa468<- subset(epa468,
car.truck=="C",select=c(mpg,vpc))
```

One-Way ANOVA: Cylinders

Now use the aov procedure to run the ANOVA.

We need to treat the variable vpc as a factor so we use the as.factor() function:

```
lm0<-aov(epa$468 ~ as.factor(vpc))
summary(lm0)
```

One-Way ANOVA: Cylinders

Now use the `aov` procedure to run the ANOVA.

We need to treat the variable `vpc` as a factor so we use the `as.factor()` function:

```
lm0<-aov(epa$468 ∼ as.factor(vpc))
summary(lm0)
```

We use Tukey's test to compare the means for 4, 6, and 8 cylinders:

```
TukeyHSD(lm0)
```

One-Way ANOVA: Cylinders

Now use the `aov` procedure to run the ANOVA.

We need to treat the variable `vpc` as a factor so we use the `as.factor()` function:

```
lm0<-aov(epa$468 ~ as.factor(vpc))
summary(lm0)
```

We use Tukey’s test to compare the means for 4, 6, and 8 cylinders:

```
TukeyHSD(lm0)
```

The results indicate that each mean is significantly different from the other two

One-Way ANOVA: Cylinders

We can estimate the actual difference in city mileage for 4, 6, and 8 cylinder cars by examining the parameter estimates from the linear model.

To compute this, enter:

```
lm0<-lm(epa$468 ~ as.factor(vpc))
summary(lm0)
```

One-Way ANOVA: Cylinders

We can estimate the actual difference in city mileage for 4, 6, and 8 cylinder cars by examining the parameter estimates from the linear model.

To compute this, enter:

```
lm0<-lm(epa$468 ~ as.factor(vpc))
summary(lm0)
```

The numbers in the `Estimate` column (not produced by the `aov` function) represents the following in terms of the parameters:

Row	Estimate	Interpretation
(Intercept)	27.7809	MPG for 4 cyls
as.factor(epa468$vpc)6	-6.3023	MPG 4 cyl - MPG 6 cyl
as.factor(epa468$vpc)8	-10.0394	MPG 4 cyl - MPG 8 cyl

One-Way ANOVA: Cylinders

We conclude that whether a car has 4, 6, or 8 cylinders makes a significant difference in the mileage.

The estimated mpg values by number of cylinders are:

Cylinders	MPG	Computed as:
4	27.78	-
6	21.48	27.78-6.30
8	17.74	27.78-10.04

Two-Way ANOVA without Interaction

Next we consider a model with two discrete predictors.

Two-Way ANOVA without Interaction

Next we consider a model with two discrete predictors.

We will then use this model to compare mileage data with two discrete factors, each with two levels:

- Factor 1: car or truck
- Factor 2: city or highway

Two-Way ANOVA without Interaction

We will generate artificial data fitting the model:

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \beta_1 X_{3i} + \beta_2 X_{4i} + e_i$$

With:

- $\mu = 5$
- $\alpha_1 = 1$
- $\alpha_2 = 5$
- $\beta_1 = 2$
- $\beta_2 = 7$
- $\sigma_e = 5$

Two-Way ANOVA without Interaction

The expected values for this model are given by the following table:

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \beta_1 X_{3i} + \beta_2 X_{4i} + e_i$$

Factor 1:

Factor 2:	Level 1	Level 2
Level 1	$\mu + \alpha_1 + \beta_1 = 5 + 1 + 2$	$\mu + \alpha_1 + \beta_2 = 5 + 1 + 7$
Level 2	$\mu + \alpha_2 + \beta_1 = 5 + 5 + 2$	$\mu + \alpha_2 + \beta_2 = 5 + 5 + 7$

Two-Way ANOVA without Interaction

Enter the following R statements:

```
mu<-5; alpha1<-1; alpha2<-5; beta1<-2;
beta2<-7
x1<-c(rep(1,100),rep(0,100));
x2<-c(rep(0,100),rep(1,100))
x3<-rep(c(rep(1,50),rep(0,50)),2)
x4<-rep(c(rep(0,50),rep(1,50)),2)
e<-rnorm(200,0,5)
class<-gl(2,50,200,labels=c("2010","2011"))
group<-gl(2,100,200,labels=c("Grp1","Grp2"))
y<-mu+alpha1*x1+alpha2*x2+beta1*x3+beta2*x4+e
art2<-data.frame(y,class,group)
```

Two-Way ANOVA without Interaction

Enter the following R statements:

```
mu<-5; alpha1<-1; alpha2<-5; beta1<-2;
beta2<-7
x1<-c(rep(1,100),rep(0,100));
x2<-c(rep(0,100),rep(1,100))
x3<-rep(c(rep(1,50),rep(0,50)),2)
x4<-rep(c(rep(0,50),rep(1,50)),2)
e<-rnorm(200,0,5)
class<-gl(2,50,200,labels=c("2010","2011"))
group<-gl(2,100,200,labels=c("Grp1","Grp2"))
y<-mu+alpha1*x1+alpha2*x2+beta1*x3+beta2*x4+e
art2<-data.frame(y,class,group)
```

We can get a boxplot of the data with:

```
boxplot(y ~ group*class)
```

Two-Way ANOVA without Interaction

We can display the means for the four cells as:

```
aggregate(art2, by=list(group,class),
FUN=mean)
```

Two-Way ANOVA without Interaction

We can display the means for the four cells as:

```
aggregate(art2, by=list(group,class),
FUN=mean)
```

Now run the ANOVA using `aov`:

```
lm0<-aov(y ~ group+class)
summary(lm0)
```

Two-Way ANOVA without Interaction

This time the ANOVA table has more rows because we have two factors in the model instead of one (hence the name "two-way analysis of variance"

Two-Way ANOVA without Interaction

This time the ANOVA table has more rows because we have two factors in the model instead of one (hence the name "two-way analysis of variance"

Row	df	Mean Sq	F Value	Pr(>F)
group	1	510.88	20.908	8.5e-06
class	1	966.04	39.535	2.0e-09
Residuals	197	24.44		

Two-Way ANOVA

Now we will run a 2 factor model (2-way ANOVA) without interaction on the EPA data using the following two factors:

- Factor 1: Car or Truck (2 levels)
- Factor 2: City or Highway (2 levels)

Two-Way ANOVA

Now we will run a 2 factor model (2-way ANOVA) without interaction on the EPA data using the following two factors:

- Factor 1: Car or Truck (2 levels)
- Factor 2: City or Highway (2 levels)

We can simplify the R code by using the `attach(epa)` statement, or we can just precede each variable name with `epa$`.

If we choose not to attach epa, the code would be:

```
lm0<-aov(epa$mpg ~ epa$C.H+epa$car.truck
summary(lm0)
```