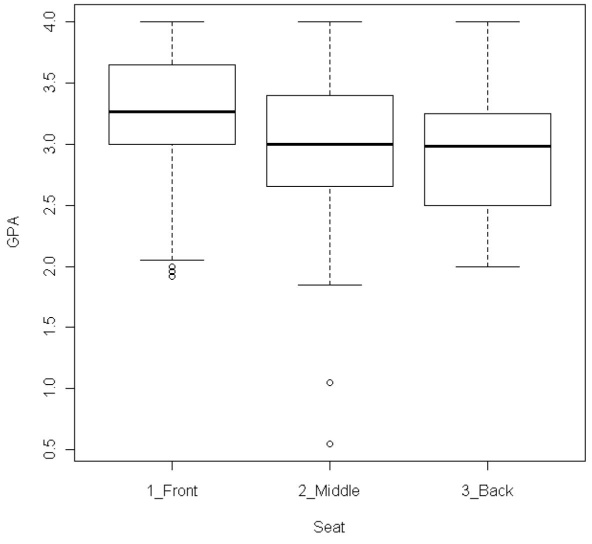
ONE-WAY ANALYSIS OF VARIANCE: GPA BY SEAT LOCATION EXAMPLE

There are 384 students in the dataset. Y = GPA and there is one categorical variable, “Seat” which is a response to the question “Where do you typically sit in a classroom – in the front, middle or back?” We want to know if population mean GPA differs for students who typically sit in the 3 classroom locations. If so, we want to know which locations have means that are significantly different.

First, here are boxplots and summary statistics of the GPAs for the three seat locations:

```
boxplot(GPA~Seat, ylab="GPA",
xlab="Seat", data=Seats)
```



```
> attach(Seats)
```

Get group means:

```
> tapply(GPA, Seat, mean)
 1_Front 2_Middle   3_Back
3.202955 2.985275 2.919359
```

Get group standard deviations:

```
> tapply(GPA, Seat, sd)
  1_Front  2_Middle    3_Back
0.5491962 0.5576177 0.5104603
```

Get group sample sizes:

```
> tapply(GPA, Seat, length)
 1_Front 2_Middle   3_Back
      88      218       78
```

Here are the commands and results for the basic ANOVA table and test. We can reject the null hypothesis that all population means are equal.

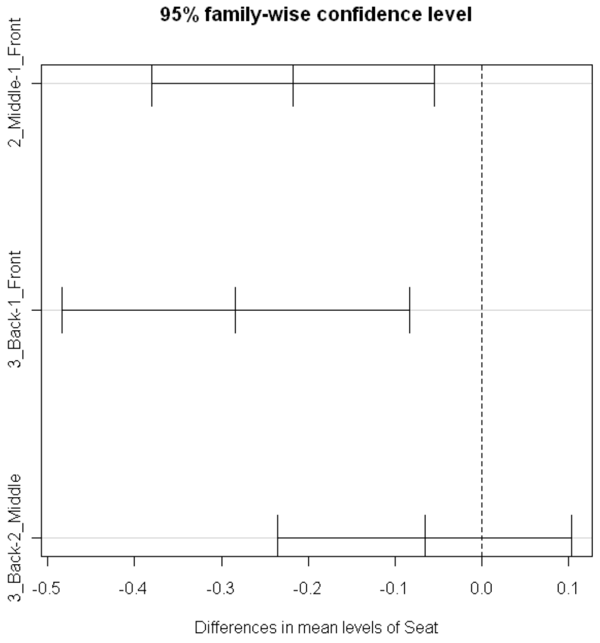
```
> Model=aov(GPA ~ Seat, data = Seats)
> summary(Model)
             Df  Sum Sq Mean Sq F value   Pr(>F)
Seat          2   3.997 1.99863  6.6927 0.001391 **
Residuals   381 113.778 0.29863
```

Here is how you get the Tukey simultaneous confidence intervals (default is overall 95% confidence). The option `ordered = T` means that you want the categories ordered according to the magnitude of the sample means. This feature is nice because it results in all of the sample differences in means being positive, making them easier to read and understand.

```
> TukeyHSD(Model, ordered = T)
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered
Fit: aov(formula = GPA ~ Seat, data = Seats)

$Seat
                       diff         lwr       upr     p adj
2_Middle-3_Back  0.06591625 -0.10372912 0.2355616 0.6316322
1_Front-3_Back   0.28359557  0.08363833 0.4835528 0.0026703
1_Front-2_Middle 0.21767932  0.05528798 0.3800707 0.0049362
```

The "p adj" provide p-values for simultaneous tests of the null hypotheses H_0: $\mu_1 = \mu_2$, etc. Here, note that the only one that *cannot* be rejected is the one comparing Middle to Back. So population mean GPAs for those two groups do not differ significantly.



The plot to the left is created using:

```
plot(TukeyHSD(Model))
```

These are Tukey confidence intervals for the differences in population means. Notice that the difference between Back and Middle covers 0, but the other two differences do not. If an interval covers 0, the difference in those two population means is not statistically significant. If the interval does not cover 0, it can be concluded that the population means for those two groups are different from each other.