MANOVA

James Jenkins

1 October 2017

MANOVA

ANOVA

In order to consider MANOVA, it is first helpful to consider the simpler case of ANOVA. ANOVA stands for "analysis of variance" and it does just that.

Specifically, all observations will exhibit some variation. If you take an abritrary factor and group the observations on that factor, the 'within group' variance will be less. Any grouping will explain some of the variance.

An ANOVA test is designed to determine whether the amount of variance explained by grouping on the variable of interest is sufficiently large to indicate some relationship. This is achieved by examining the ratio of the *within group* variance and the *between group* variance. This is the *F statistic*. We know ahead of time how much of the variance we would expect a random grouping to explain: it is governed by the *F distribution*. By looking up the value of the *F distribution*, we can determine the critical value of F below which the explained variance is not significant.

A one-way ANOVA with only two levels is a t-test.

ANOVA tests the null hypothesis:

$$\text{null:} \mu_1 = \mu_2 = ... = \mu_n$$

The *F statistic* is calculated as:

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} F = \frac{SS_{\text{within group}}/(M-1)}{SS_{\text{between group}}/(n-M)}$$

where SS indicates the sum of squares, M is the number of groups, n is the number of observations.

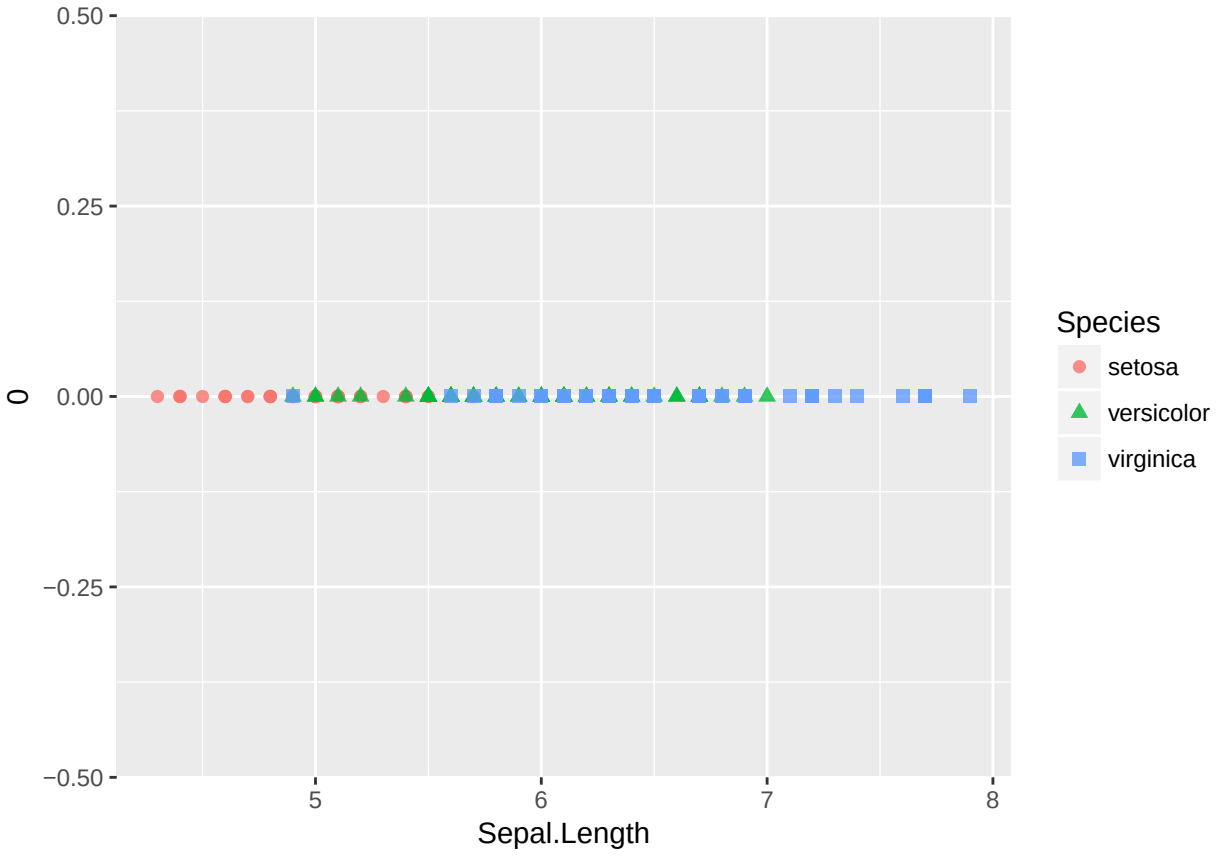
The null hypothesis is rejected if $F > F_{\text{critical}}$ where F_{critical} depends on the number of degrees of freedom and the required significance (α).

ANOVA example using `iris` dataset

```
df <- iris
head(df)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
ggplot()+
  geom_point(data = iris,
             aes(x = Sepal.Length, y = 0, colour = Species, shape = Species),
             size = 2,
             alpha = 0.8)
```



For example, we can look at the `iris` dataset that comes with `base` R. It shows the values, in cm, of sepal length and width and petal length and width for 50 flowers from each of 3 species of iris. What if we wished to discover whether the sepal length were significantly different for any of the three different species? This is plotted out above.

```
df %>%
  group_by(Species) %>%
  mutate(avg.sepal.length   = mean(Sepal.Length),
         resid.sepal.length = Sepal.Length - avg.sepal.length,
         within.group       = resid.sepal.length^2) %>%
  ungroup() %>%
  mutate(global.mean.sepal.length = mean(Sepal.Length),
         group.resid.sepal.length = global.mean.sepal.length - avg.sepal.length,
         between.group            = group.resid.sepal.length^2) -> df

ss <- data.frame(within.group  = sum(df$within.group),
                 between.group = sum(df$between.group))

ss
```

```
##   within.group between.group
```

```
## 1      38.9562      63.21213
```

The above formula first groups by the `Species` field and calculates the mean `Sepal.Length`. Then, for each `Species`, the residuals of each of the observations on that `Species` are taken from the `Species` mean and squared. This is *within group* or residual deviation.

The second part ungroups the data and, for each observation, compares the `Species` mean to the global mean. The squared residuals are taken as the *between group*.

If, hypothetically, you had one observation at each group, the group mean would equal the observation value such tha tthe *within group* variation would be zero. The *between group* variation would then explain all of the variation because each group mean would perfectly describe the corresponding observation.

The *between group* variation is the amount of variation explained by having different averages, whereas the *within group* variation is the amount of variation not explained by having different averages and why there is still some variation in each group around the mean.

```
dof_between <- length(levels(df$Species))-1
dof_within <- nrow(df)-(dof_between+1)

f_stat <- (ss$between.group/dof_between)/
  (ss$within.group/dof_within)
f_stat
```

```
## [1] 119.2645
```

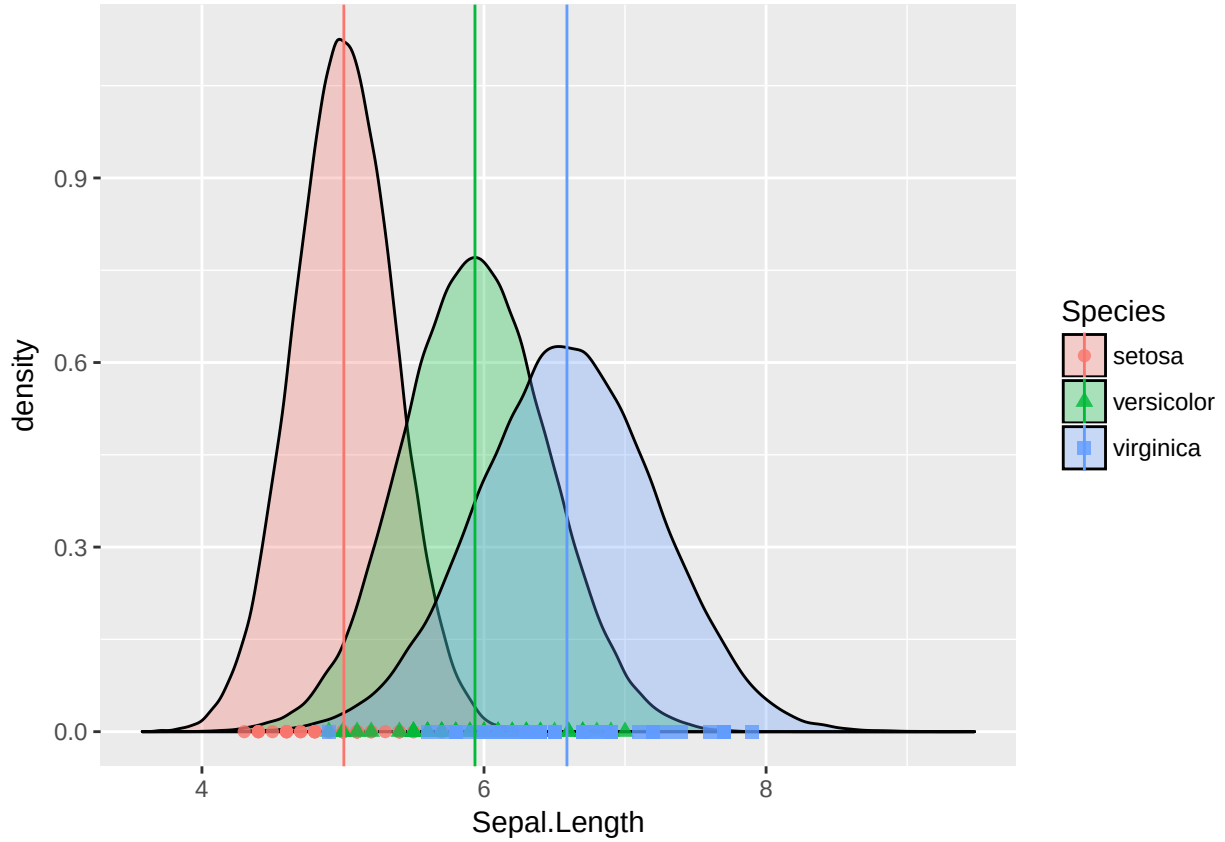
The obove calculates the F-statistic for the test. The number of degrees of freedom assigned to within and between groups are 2 and 147 respectively. Since the critical value of the F distribution at 95% confidence is `qf(.95, 2, 147, lower.tail = FALSE)`$= 0.0513112$, the null hypothesis is rejected. Through trial of ever decreasing confidence levels, it can be determined that the p-value is negligable.

Of course, all of this calculation can be achieved through a simple command in R:

```
summary(aov(Sepal.Length ~ Species, data = iris))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## No id variables; using all as measure variables
```



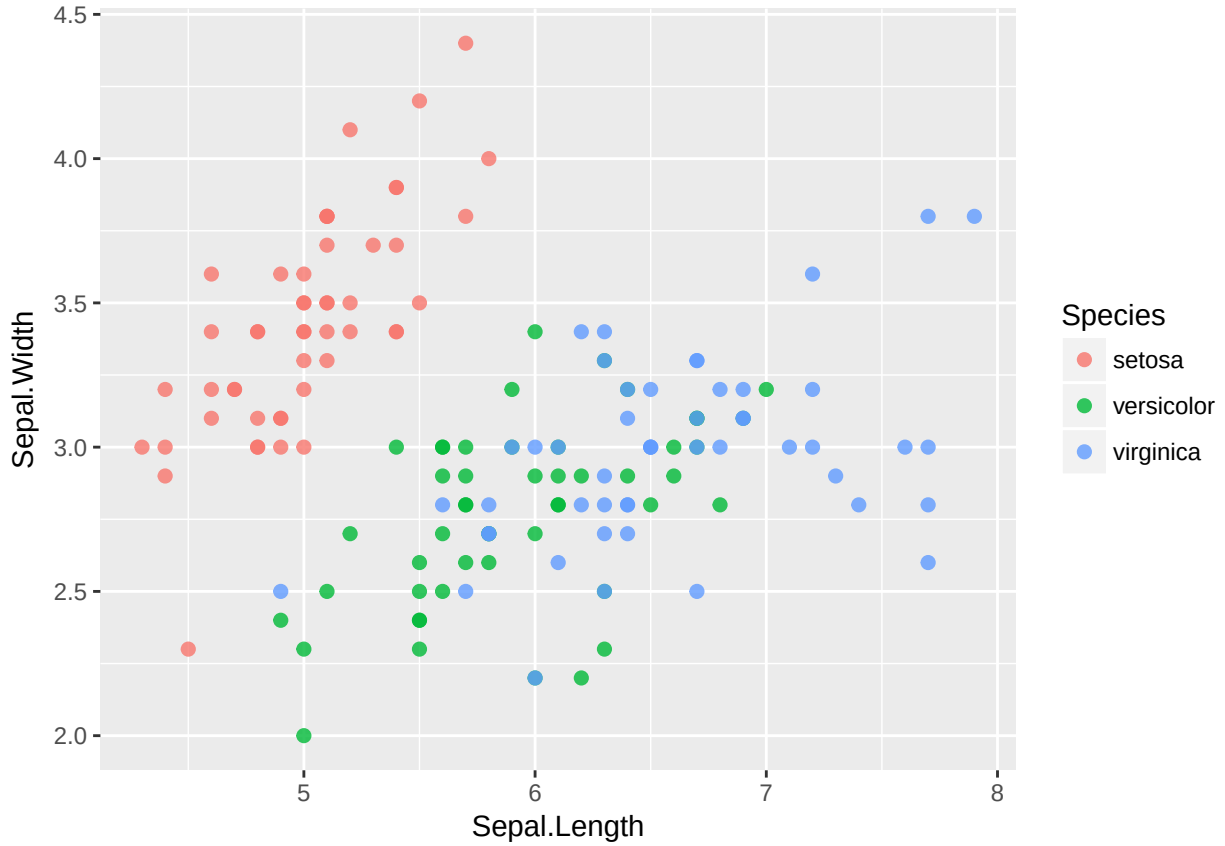
If we plot the observed Sepal Lengths along the x-axis and superimpose the distributions of the three groups, we obtain a visual representation of the test. Although the three groups have considerable overlap, it looks obvious that the setosa and virginica species have qutie different means. In truth, versicolor and virginicia also have significantly differnt means.

MANOVA

MANOVA is the extension of ANOVA to more than one dimension - i.e. it allows for the testing of multiple dependent variables at the same time. For example, whether there is a significant difference in some linear combination of Sepal Length and Sepal Width between Species?

MANOVA example using `iris` dataset

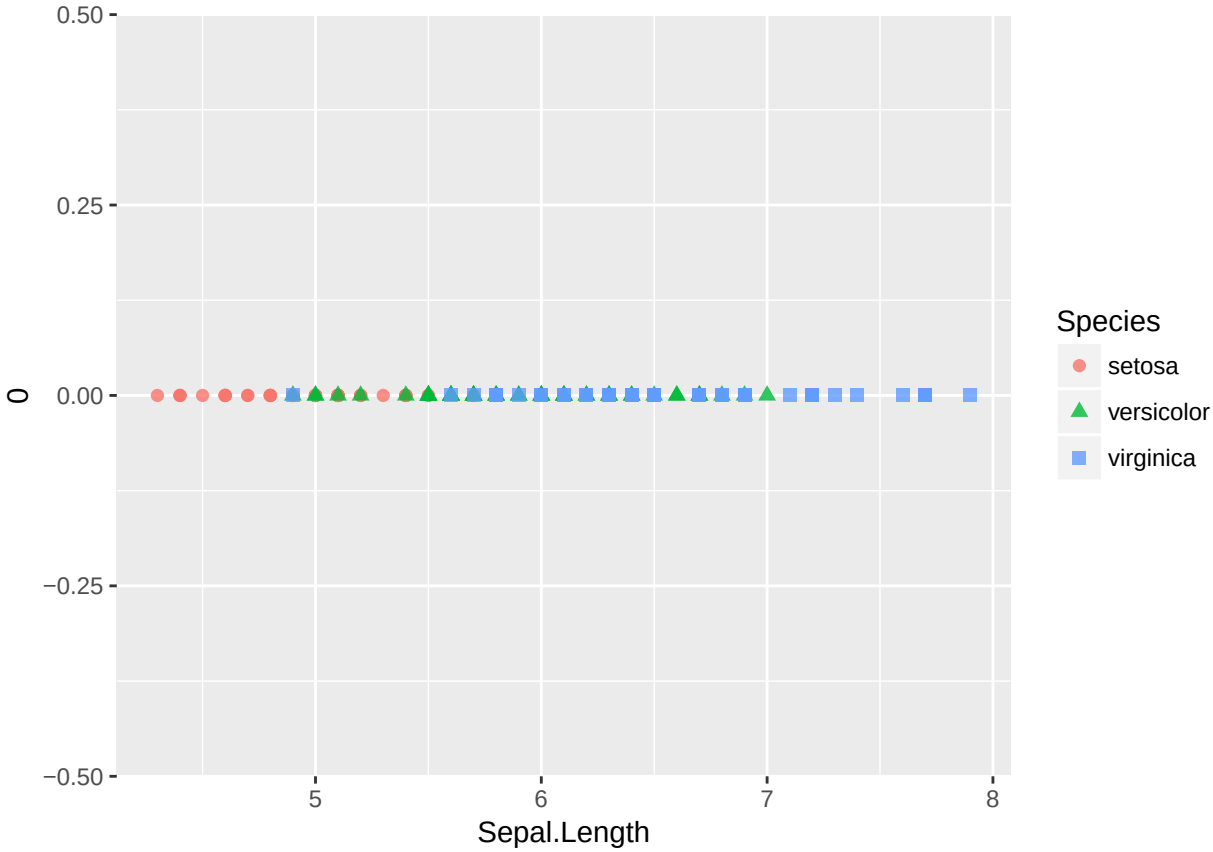
```
ggplot(iris,
       aes(x = Sepal.Length, y = Sepal.Width, colour = Species, Shape = Species))+
  geom_point(size = 2, alpha = 0.8)
```



Now, we already know from our prior ANOVA that the three groups are separate on the Sepal.Length axis but we do not know whether they are significantly different on the Sepal.Width axis. Moreover, had we not done the first ANOVA test, it might not be immediately obvious that the three groups are significantly different.

Imagine we were presented with the three groups and their 300 observations (50 per group per dependent variable). We could approach the problem by running a series of ANOVA tests, one for each axis. In this case we would first run ANOVA on the x-axis, Sepal.Length:

```
ggplot()+
  geom_point(data = iris,
             aes(x = Sepal.Length, y = 0, colour = Species, shape = Species),
             size = 2,
             alpha = 0.8)
```

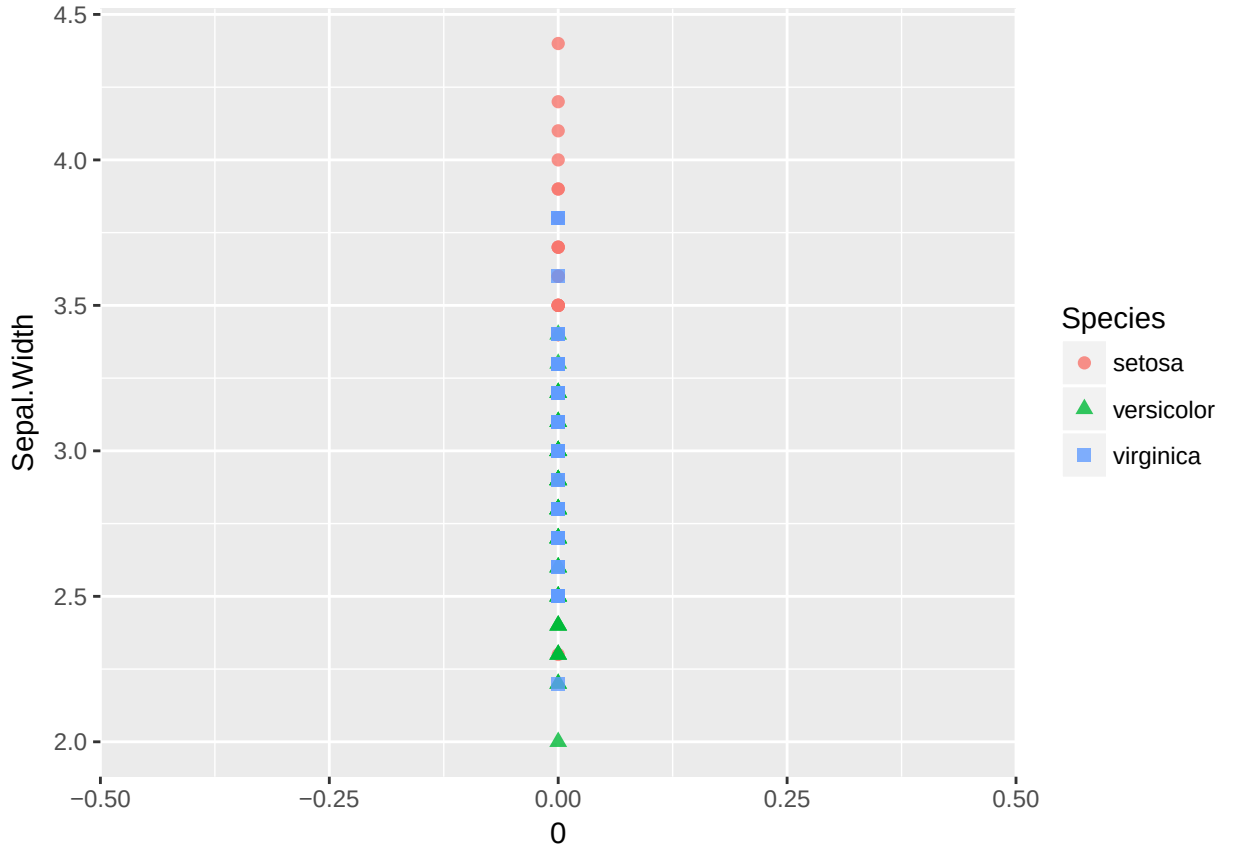



```
summary(aov(Sepal.Length ~ Species, data = iris))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then we would run a second ANOVA an the y-axis, Sepal.Width:

```
ggplot()+
  geom_point(data = iris,
             aes(x = 0, y = Sepal.Width, colour = Species, shape = Species),
             size = 2,
             alpha = 0.8)
```




```
summary(aov(Sepal.Width~ Species, data = iris))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  11.35   5.672   49.16 <2e-16 ***
## Residuals   147  16.96   0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In each case, we are taking the projection of the data onto the relevant axis and testing whether the means are significantly different on that axis only.

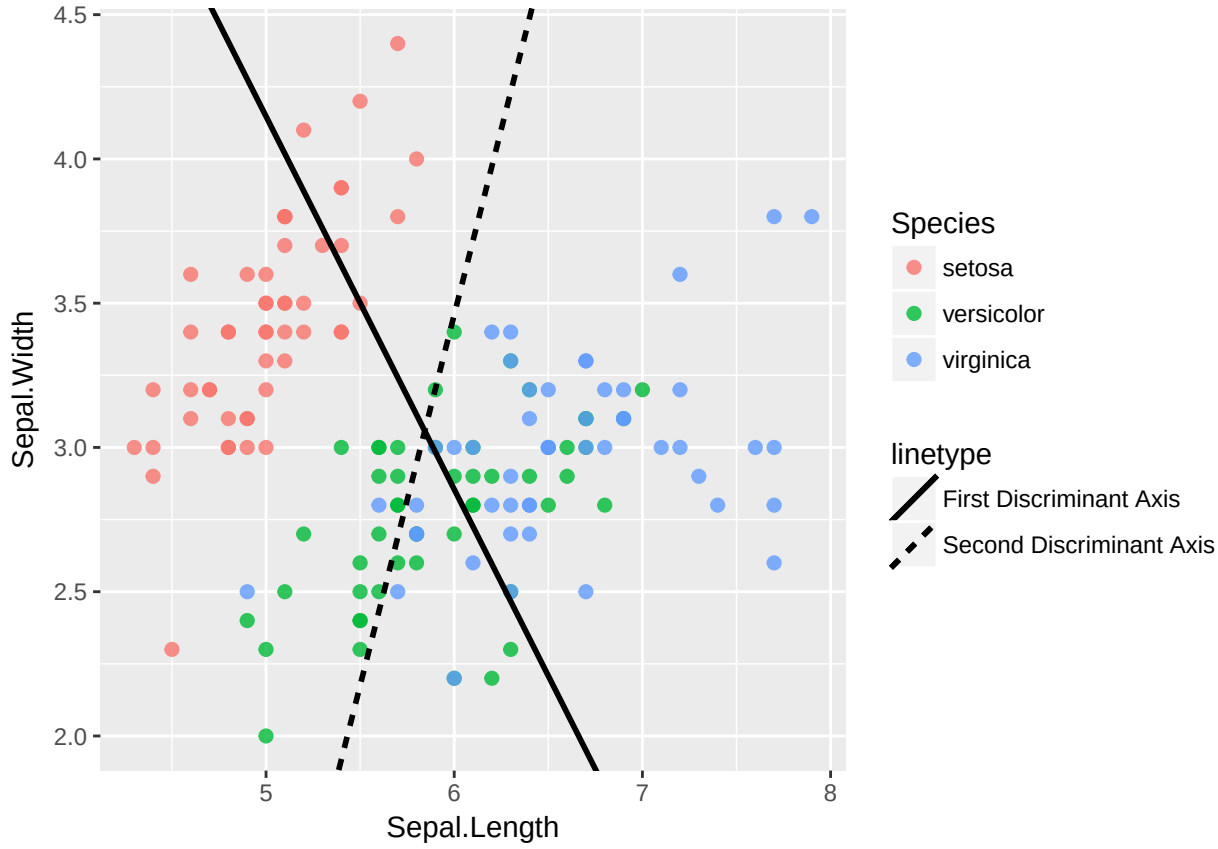
There are numerous experimental designs in which this may be the desired approach. In some cases, it may not make sense to combine your dependent variables linearly or you might expect very little correlation between the dependent variables, in which case MANOVA has little to offer over sequential ANOVA.

If, however, it is reasonable to consider a linear combination of your dependent variables, then a MANOVA offers a truly multivariate approach.

```
discriminant <- lda(Species ~ Sepal.Length + Sepal.Width, data = iris)

global_mean_length <- mean(iris$Sepal.Length)
global_mean_width <- mean(iris$Sepal.Width)
grad_lda1 <- discriminant$scaling["Sepal.Width", "LD1"]/
  discriminant$scaling["Sepal.Length", "LD1"]
intercept_lda1 <- global_mean_width - global_mean_length*grad_lda1
grad_lda2 <- discriminant$scaling["Sepal.Width", "LD2"]/
  discriminant$scaling["Sepal.Length", "LD2"]
intercept_lda2 <- global_mean_width - global_mean_length*grad_lda2
```

```
ggplot(iris,
       aes(x = Sepal.Length, y = Sepal.Width, colour = Species, Shape = Species))+
  geom_point(size = 2,
             alpha = 0.8)+
  geom_abline(aes(intercept = intercept_lda1, slope = grad_lda1, linetype = "First Discriminant Axis"),
              size = 1)+
  geom_abline(aes(intercept = intercept_lda2, slope = grad_lda2, linetype = "Second Discriminant Axis")
              size = 1)
```

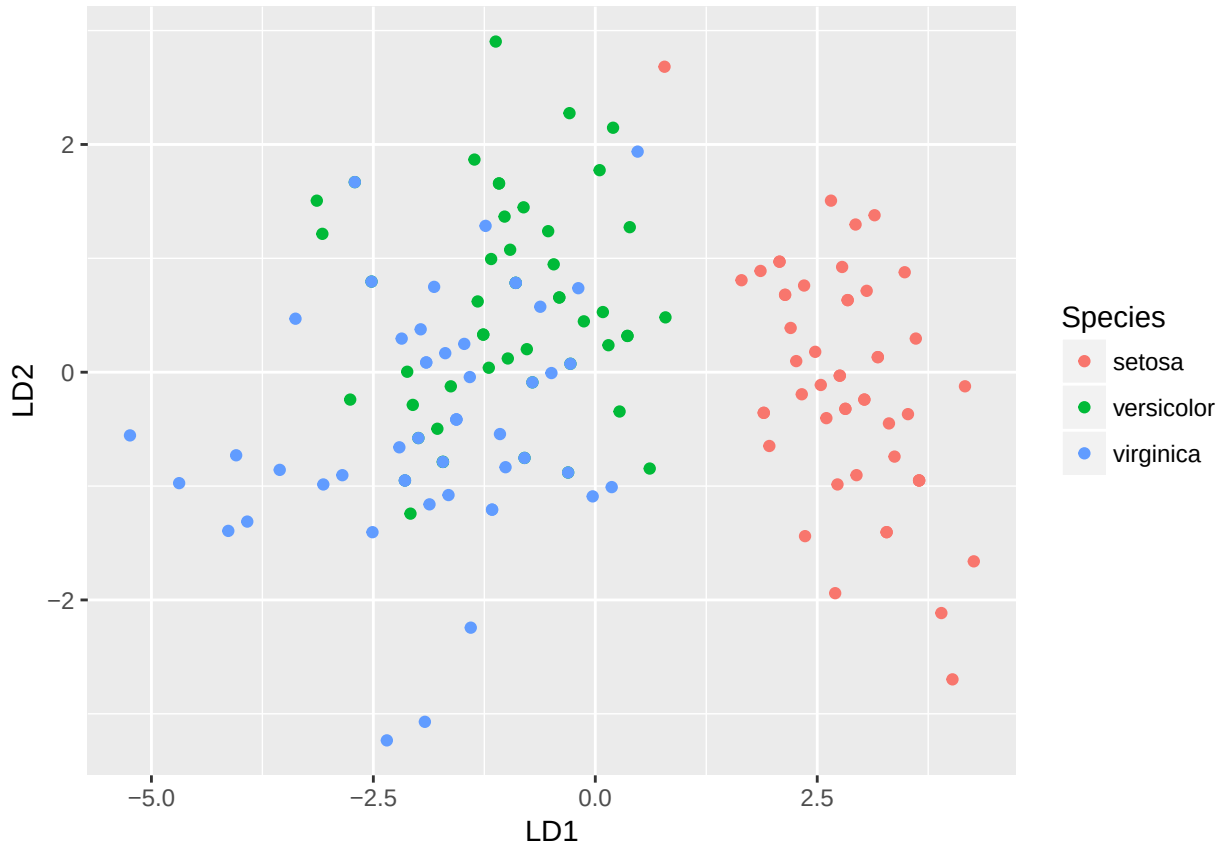


MANOVA is related to linear discriminant analysis (LDA) (Borgen and Seling 1978). It tests whether there exists *any* linear combination of dependent variables (e.g. Sepal.Width, Sepal.Length) over which there is a significant difference between the groups. Intuitively, the axis that maximally separates the groups is the most likely to have a significant difference.

The plot above shows the first two discriminant axes of the iris dataset over the two Sepal variables. The first axis is the solid line which is almost orthogonal to the major axis of the setosa group and the combined group of versicolor and verginica. When these groups are then projected onto the first and second axes, they are maximally separated.

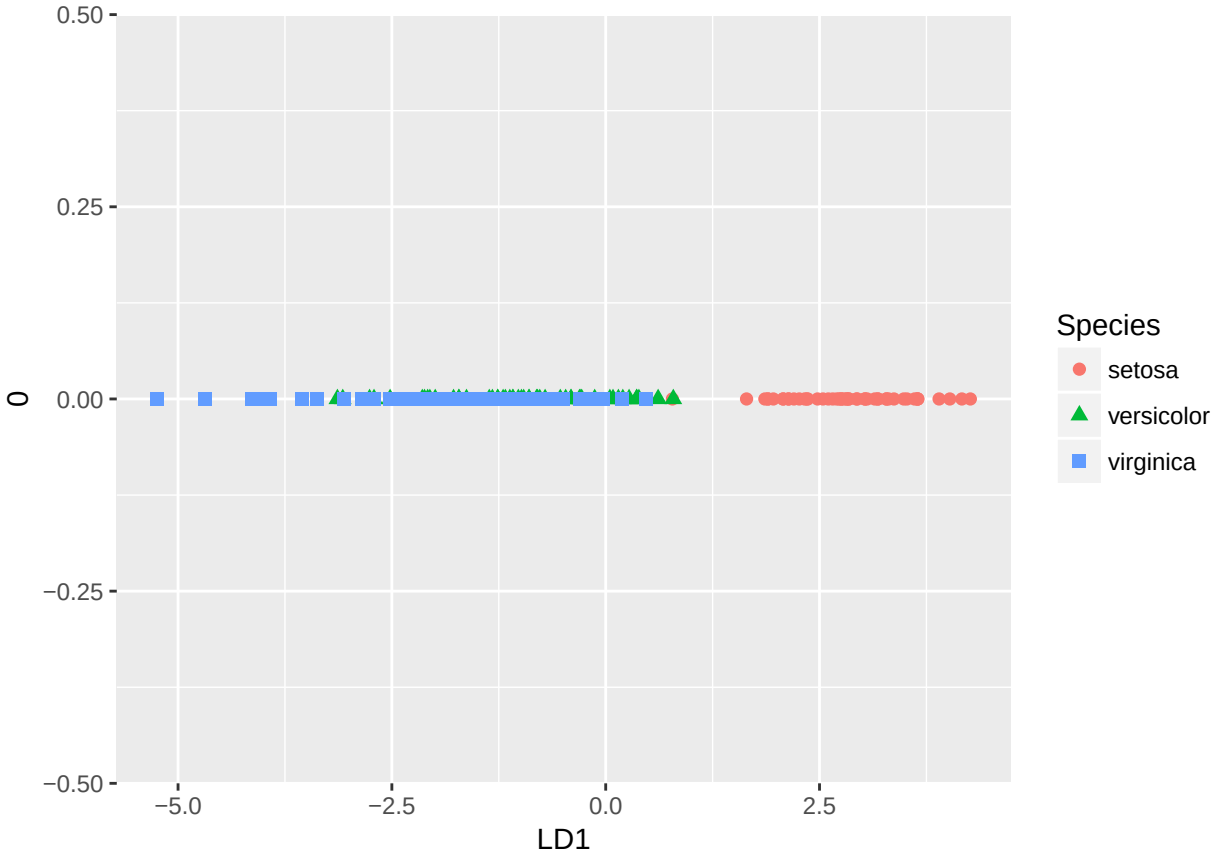
```
discriminant_prediction <- data.frame(Species=iris$Species,predict(discriminant)$x)

ggplot(discriminant_prediction,
       aes(x = LD1, y = LD2, colour = Species))+
  geom_point()
```



The above plot shows the iris dataset plotted against LD1 and LD2 instead of Sepal.Length and Sepal.Width.

```
ggplot(discriminant_prediction,
       aes(x = LD1, y = 0, colour = Species))+
  geom_point(aes(shape = Species), size = 2)
```



The graph above shows the projection of the iris dataset onto just the first discriminant axis and clearly shows that the setosa group has been separated much more than it is on either the Sepal.Length or Sepal.Width axes.

Now, having performed this lda we can perform a MANOVA by using the simple command in R:

```
man1 <- manova(cbind(Sepal.Length, Sepal.Width)~Species, iris)
summary(man1)
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## Species     2 0.94531   65.878      4    294 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the sake of completeness, we can use the `car` package to perform a MANOVA with Roy's statistic:

```
man2 <- Manova(lm(cbind(Sepal.Length, Sepal.Width)~Species, iris), test = "Roy")
man2
```

```
##
## Type II MANOVA Tests: Roy test statistic
##         Df test stat approx F num Df den Df    Pr(>F)
## Species  2    4.1718   306.63      2    147 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

which gives a Roy's lambda of 4.1718. This corresponds to a θ value of 0.8066437. I calculate this so that I can replicate the calculations perform by Grice and Iwasaki (Grice and Iwasaki 2007).

```r
summary(aov(discriminant_prediction$LD1~iris$Species))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## iris$Species   2  613.3   306.6   306.6 <2e-16 ***
## Residuals    147  147.0     1.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running an ANOVA on the data projected onto the first discriminant axis, we obtain an F value of 306.6. Following Grice and Iwasaki's methodology we can see that:

$$\frac{F_{observed} df_{between}}{F_{o}bserved df_{b}etween + df_{within}} = \frac{306.6 \times 2}{306.6 \times 2 + 147}$$

This gives a value of 0.8066298, which is equal to Roy's θ above.

References

Borgen, Fred H, and Mark J Seling. 1978. "Uses of Discriminant Analysis Following Manova: Multivariate Statistics for Multivariate Purposes." *Journal of Applied Psychology* 63 (6). American Psychological Association: 689.

Grice, James W, and Michiko Iwasaki. 2007. "A Truly Multivariate Approach to Manova." *Applied Multivariate Research* 12 (3): 199–226.