ANOVA_script

Kevin Healy

2023-10-23

In this script we will look at using ANOVA in R.

Data

We will stick with using the iris dataset since we are familiar with it and its already uploaded.

```
iris_data <- iris
```

We will also use violin plots so lets upload the vioplot package which you installed in previous weeks.

```
library(vioplot)
```

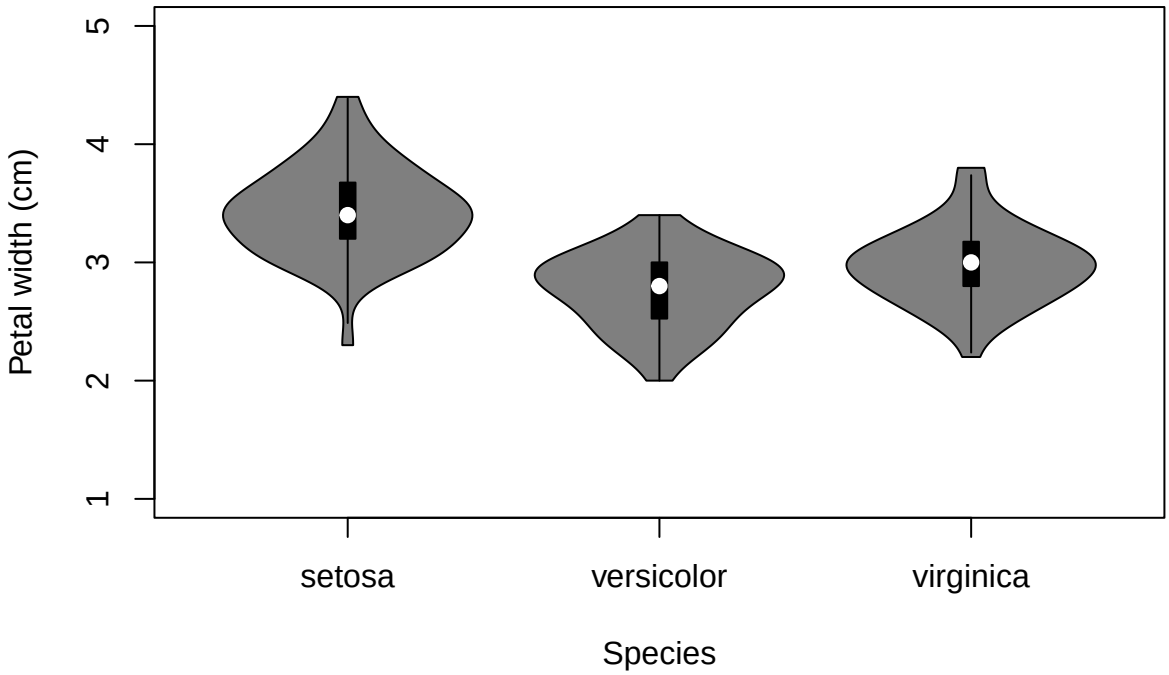
```
## Loading required package: sm

## Package 'sm', version 2.2-5.7: type help(sm) for summary information

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

ANOVA

Last week we looked at compared two iris species petal and sepal morphology. However, what if we wanted to simply ask if the petal width was different between the 3 species in our dataset. If we used t-test we would have to run 3 t-test (setosa vs versicolor; setosa vs virginica; versicolor vs virginica). The problem with this is when you run multiple tests you increase the chance of a type II error (saying something is significantly different when the difference was really just down to chance). Remember how running our coin flipping experiment enough times eventual gave a significant value saying the coin was biased even though we know it was fair. This is the same problem with running lots of t-test and is called multiple testing. Hence, when possible, we want to run a single test comparing all groups using the ANOVA.

Lets compare the petal width between all three species. First lets plot this so we can get a feel of what the data looks like. We will use a violin plot as its helpful to visually see the distributions of data for each species.

```
vioplot(iris$Sepal.Width ~ iris$Species,
     xlab = "Species",
     ylab = "Petal width (cm)",
     ylim = c(1,5))
```



From the plot it looks like there is a difference in petal width between each of the species. However, lets use an ANOVA to test this with the 'aov()' function.

Notice that when we use the ANOVA we will use the formula Y ~ X, with the response variable on the Y axis (Petal.Width) and explanatory variable (Species) on the x-axis. We will save the model as an object called mod1 and then use the 'summary()' function to look at the results.

```
#run the ANOVA
mod1 <- aov(iris$Petal.Width ~ iris$Species)

#use summary() to look at the results
summary(mod1)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## iris$Species   2  80.41   40.21     960 <2e-16 ***
## Residuals    147   6.16    0.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the summary results section we can notice 2 rows and 5 columns. The first row represents the between group variance, which in our case are the different iris species. This is the how far each of the groups is from the overall mean. The first column DF is the degrees of freedom for this estimate, which for the within group estimate is the number of groups - 1. The Sum Sq is the sum of the squares which is our estimate of variance (think back at how var() is measured in lecture 3). The Mean Sq is just the Sum Sq divided by the Df to account for sample size and will be used to calculate the F value.

The second row is the residuals. This is the within group variance or how far each of the data points was from its groups mean. Notice the degrees of freedom for within group variance is much higher as its the full

samples size 150 minus the number of groups 3. The Sum Sq and Mean Sq are calculated in a similar way to above.

The F value is then calculated as the Mean Sq of between variance divided by the Mean Sq of within variance. This F value is then compared to a table to calculate a p-value relating to that value being greater than 1, that is the probability that the F value is greater than 1 when in reality it is not.

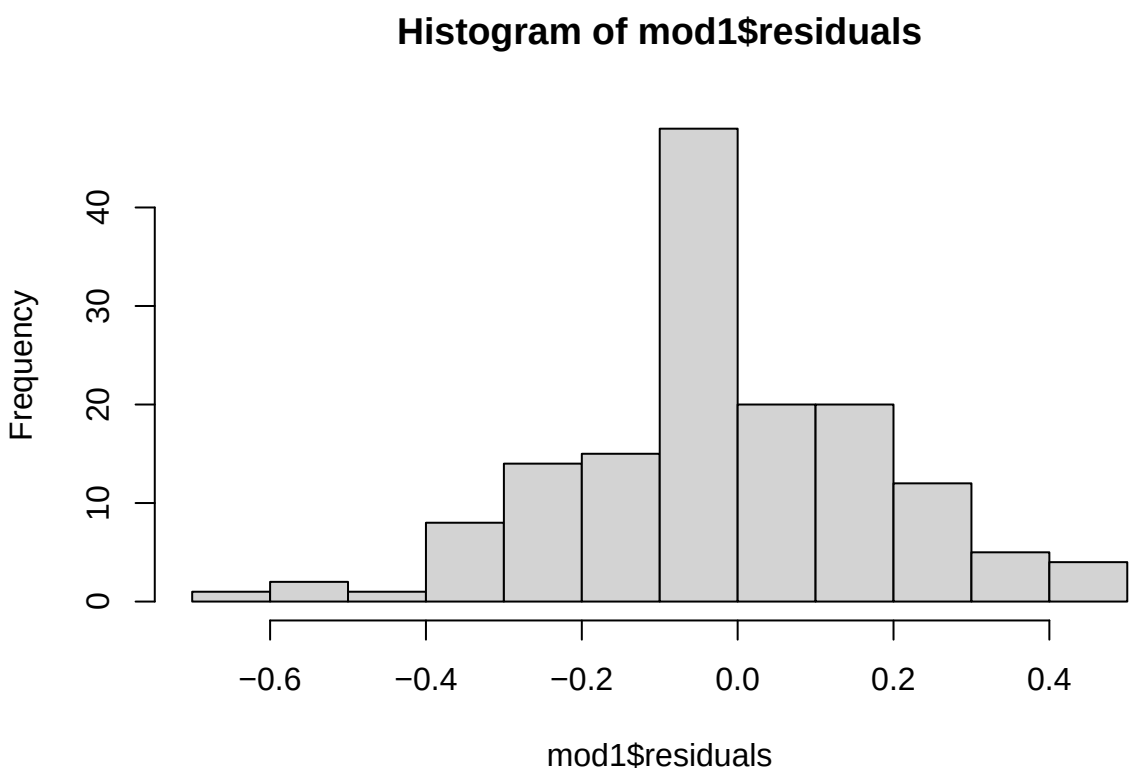
In this case we can reject the Null hypothesis (that there is no difference in petal length in species). In a results section we would write something like "A one-way ANOVA showed that there was a significant difference between petal width in the three species of Iris we tested ($F(2,147) = 960$, $p < 0.05$)".

Here the statistics given in the brackets are (F(df between groups, df within groups) = F-value, $p < 0.05$).

checking the model

When running a model you should always check its assumptions. One important set of assumptions is that the residuals are normal, that is the distances of the data points from the fitted value (which is just the group means in this case) are normally distributed. We can check this by plotting the residuals directly using a histogram.

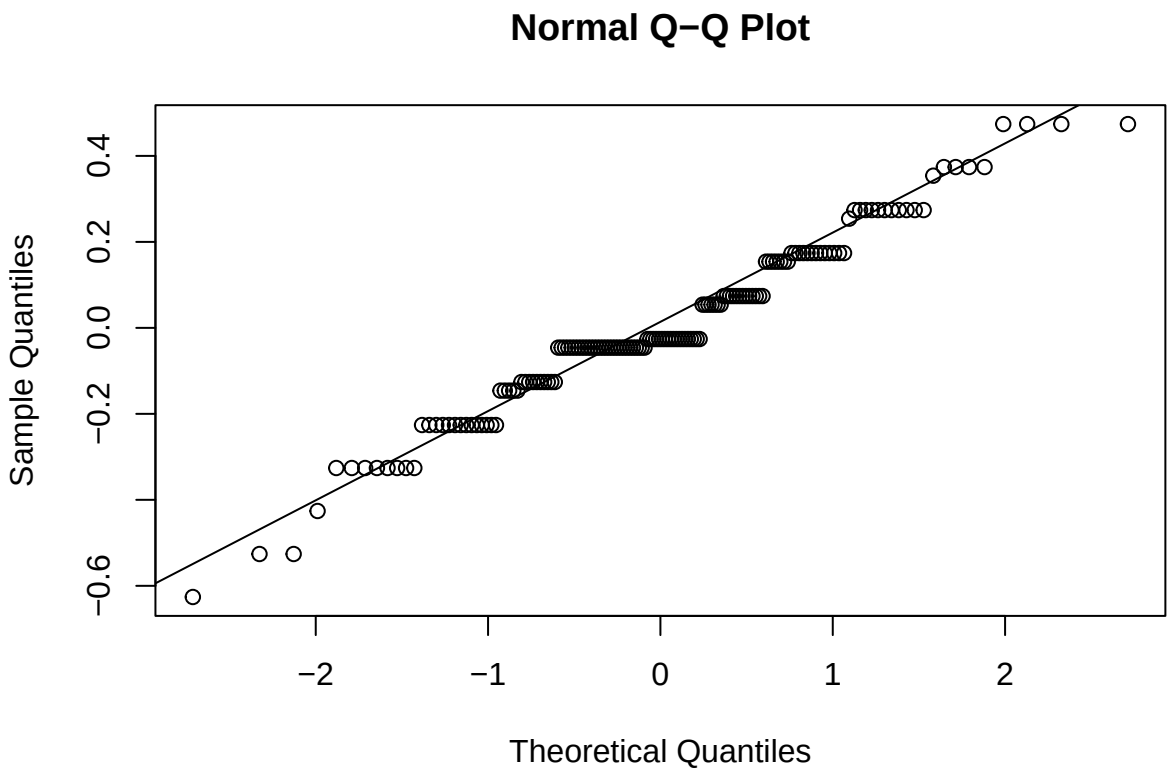
```
hist(mod1$residuals)
```



However, it can difficult to decide if this distribution is normal enough. One approach is to use a QQ-plot. This is were we compare the distribution of our residuals to an ideal normal distribution. Lets plot it using the below code.

```
#plot a qq plot for normal data
qqnorm(mod1$residuals)
```

```
#Add the line for a normal distribution
qqline(mod1$residuals)
```



The QQ-plot has Theoretical values on the x-axis and our residual values on the y axis. The graph compares how close our residuals are to where they would be expected for an idealized normal distribution which is given by the dotted line. The close the residuals to the dotted line the closer to a normal distribution they are.

Here the residuals follow the line quite well so this looks good. Like many aspects of statistics, this is just a tool to help you decide if your model is working the way you want it. This means deciding on how good is good enough is subjective and more of an art.

Post hoc testing

Notice how the null hypothesis when using an ANOVA is that there is no difference between groups with the alternative hypothesis just that there is a difference. This means we don't test which group is different. If we want to test this we need to do what is called a post hoc test, which will compare each of the groups. We will use Tukey's test which is similar to a t-test but which simultaneously compares all groups in a pairwise manner and uses the range of difference in the means. While its a useful test it does require the sample size in each group to be similar (balanced design)

```
#Run the Tukey test
tukey_res <- TukeyHSD(mod1)

#some test don't require using summary() this is one of them
tukey_res
```

```
##   Tukey multiple comparisons of means
```

```
##     95% family-wise confidence level
##
## Fit: aov(formula = iris$Petal.Width ~ iris$Species)
##
## $`iris$Species`
##                        diff       lwr       upr p adj
## versicolor-setosa    1.08 0.9830903 1.1769097     0
## virginica-setosa     1.78 1.6830903 1.8769097     0
## virginica-versicolor 0.70 0.6030903 0.7969097     0
```

In the output we are given each of the pairwise comparisons, the difference between the means (diff), the lwr and upr are the lower and upper critical values which are calculated from a table of Studentized range values (we won't worry about that here). What we can see across all pairwise comparisons is that they are significantly different.

Writing this in a results section would look something like "A one-way ANOVA showed that there was a significant difference between petal width in the three species of Iris we tested ($F(2,147) = 960$, $p < 0.05$). In a Tukey's HSD post hoc test we found that each of the pairwise comparisons where significant, with a mean difference in petal width of 1.08 between versicolor and setosa ($p < 0.5$), 1.78 between virginica and setosa ($p < 0.5$) and 0.7 between virginica and versicolor ($p < 0.5$)."

Kruskal–Wallis test

If our data and resulting residuals are not normally distributed we might want to use a non-parametric version of the test. The Kruskal–Wallis extends on the Mann–Whitney U test by comparing the ranks of each of the groups instead of the means and variances. It only assumes each of the distributions are similar.

Lets run the above test using a Kruskal–Wallis test.

```
#run the ANOVA
kw_mod <- kruskal.test(iris$Petal.Width ~ iris$Species)

#we don't need a summary() function for this function.
kw_mod
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  iris$Petal.Width by iris$Species
## Kruskal-Wallis chi-squared = 131.19, df = 2, p-value < 2.2e-16
```

Unlike the ANOVA we get a simple output here, the chi-squared value which is used to calulate the p value, the degree of freedom for the number of groups and the p value, which here is clearly much lower than 0.05.

Like above we might want to test which groups are different. To do this we can't use a Tukey test as our data is not normal but we can use a series of Mann– Whitney U tests with he p-values adjusted to account for the inflated probability of a type I error.

```
#Notice that the response variable and explanatory variable are wrote
#differently to the Y ~ X. This is just a quirk of this particular function.

pairwise.wilcox.test(iris$Petal.Width, iris$Species,
                     p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  iris$Petal.Width and iris$Species
##
##            setosa  versicolor
## versicolor < 2e-16 -
## virginica  < 2e-16 2.9e-16
##
## P value adjustment method: bonferroni
```

The results are given in a little matrix, with the p-value for each pairwise given according to row and column. For example, the p-value for the comparison between virginica and versicolor is 3e-16. All the values are clearly lower than ($p < 0.05$).