Lab 7: Cuckoos

Cori Carver

March 11 2021

General information

This lab is due October 28th by 11:59 pm. This lab is worth 10 points (each question is worth 1 point unless otherwise noted). You must upload your .rmd file and knitted PDF to Canvas. You are welcome and encouraged to talk with classmates and ask for help. However, each student should turn in their own lab assignment and all answers, including all code, needs to be solely your own.

Objective

The goal of this lab is to run and interpret ANOVA tests, including testing whether assumptions are met and visually interpreting data.

Background

The European cuckoo does not look after its own eggs, but instead lays them in the nests of birds of other species. This is known as *brood parasitism*. It has been documented previously that cuckoos have evolved to lay eggs that are colored similarly to the host birds' eggs. Is the same true of size? Do cuckoos lay eggs of different sizes in nests of different hosts? We will investigate this question, using the data file "cuckooeggs.csv". This file contains data on the lengths of cuckoo eggs laid in a variety of other species' nests.

Exploring the data and testing assumptions

First, read in the datafile "cuckooeggs.csv" and take a look at the data.

Question 1 Look at the structure of the cuckoo data. What is the explanatory variable? What is the response variable?

```r
cuckoos <- read.csv("cuckooeggs.csv")
str(cuckoos)
```

```
## 'data.frame':    120 obs. of  2 variables:
##  $ Host.Species: chr  "Hedge Sparrow" "Hedge Sparrow" "Hedge Sparrow" "Hedge Sparrow" ...
##  $ Egg.Length  : num  20.9 21.6 22.1 22.9 23.1 ...
```

```r
# code to convert the host-species variable to a "factor", i.e. categorical data
cuckoos$Host.Species = as.factor(cuckoos$Host.Species)
```

The explanatory variable is the host species and the response variable is the egg length.

Question 2 How many species of birds were measured in this study? Using the ‘str‘ function (on the entire dataframe object) or ‘levels‘ function (on the column of interest) is an easy way to check this.

```
str(cuckoos)
```

```
## 'data.frame':    120 obs. of  2 variables:
##  $ Host.Species: Factor w/ 6 levels "Hedge Sparrow",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Egg.Length  : num  20.9 21.6 22.1 22.9 23.1 ...
```

There were 6 different bird species measured in this study.

To test whether the data are distributed normally *within each group*, use a Shapiro-Wilk Normality Test. You'll need to run a test for each species, meaning you'll want to use the subset function to break up your dataset by species. Here is one example to get you going…

```
cuckoo_HS<-subset(cuckoos, cuckoos$Host.Species=="Hedge Sparrow")

shapiro.test(cuckoo_HS$Egg.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cuckoo_HS$Egg.Length
## W = 0.94843, p-value = 0.5366
```

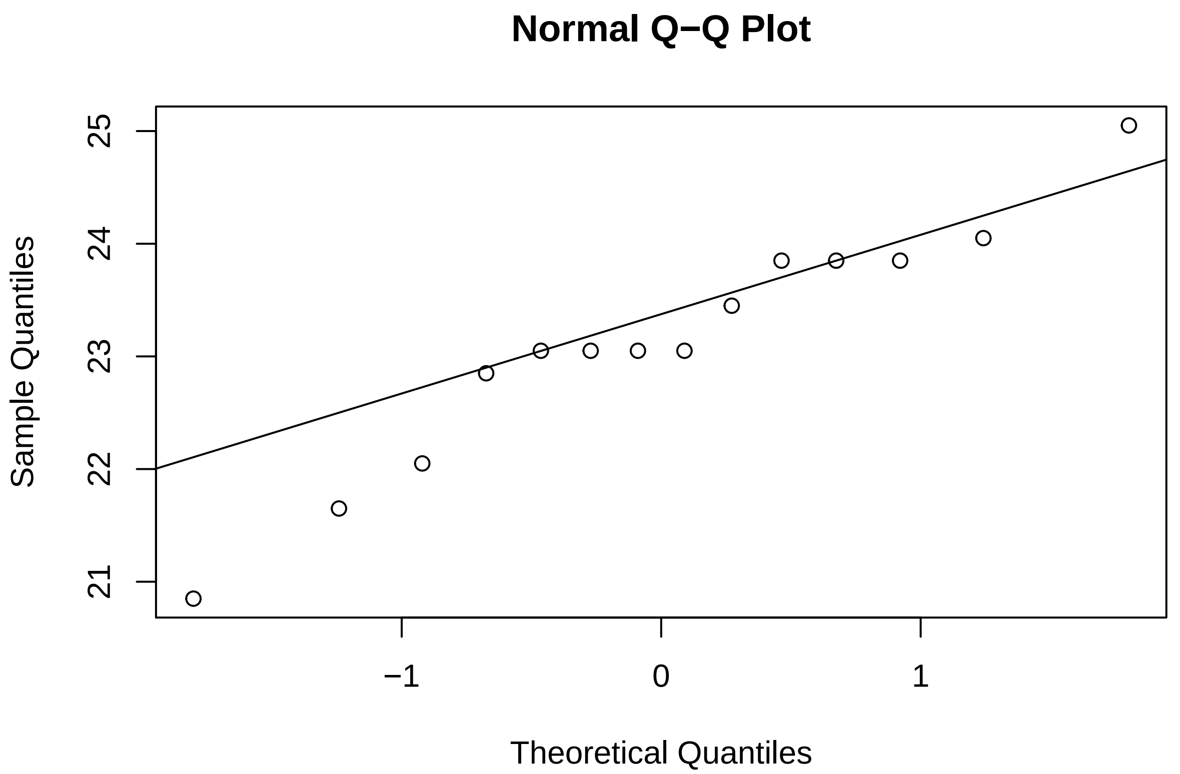
```
# Or

shapiro.test(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Hedge Sparrow"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$Egg.Length, cuckoos$Host.Species == "Hedge Sparrow")
## W = 0.94843, p-value = 0.5366
```

Since we haven't used qq-plots in lab yet, below is an example. Tip: recall from earlier labs, you can put multiple commands for drawing an individual plot inside curly braces "{ }" to run as a chunk.

```
{qqnorm(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Hedge Sparrow"))
qqline(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Hedge Sparrow"))}
```

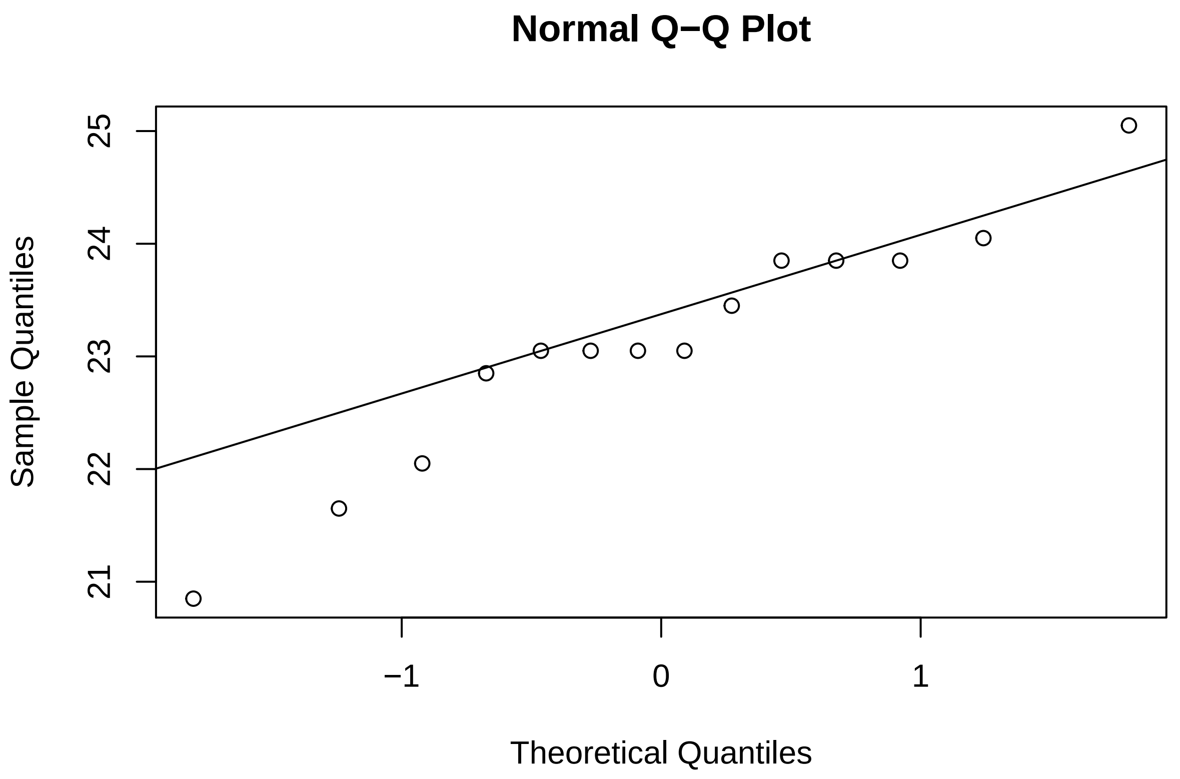


Question 3 Using the Shapiro-Wilk test for normality, as well as visual inspection of the data (i.e., plotting), evaluate whether cuckoo egg length data is normally distributed *within each group*. Interpret the output of the test.

```
shapiro.test(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Hedge Sparrow"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$Egg.Length, cuckoos$Host.Species == "Hedge Sparrow")
## W = 0.94843, p-value = 0.5366
```

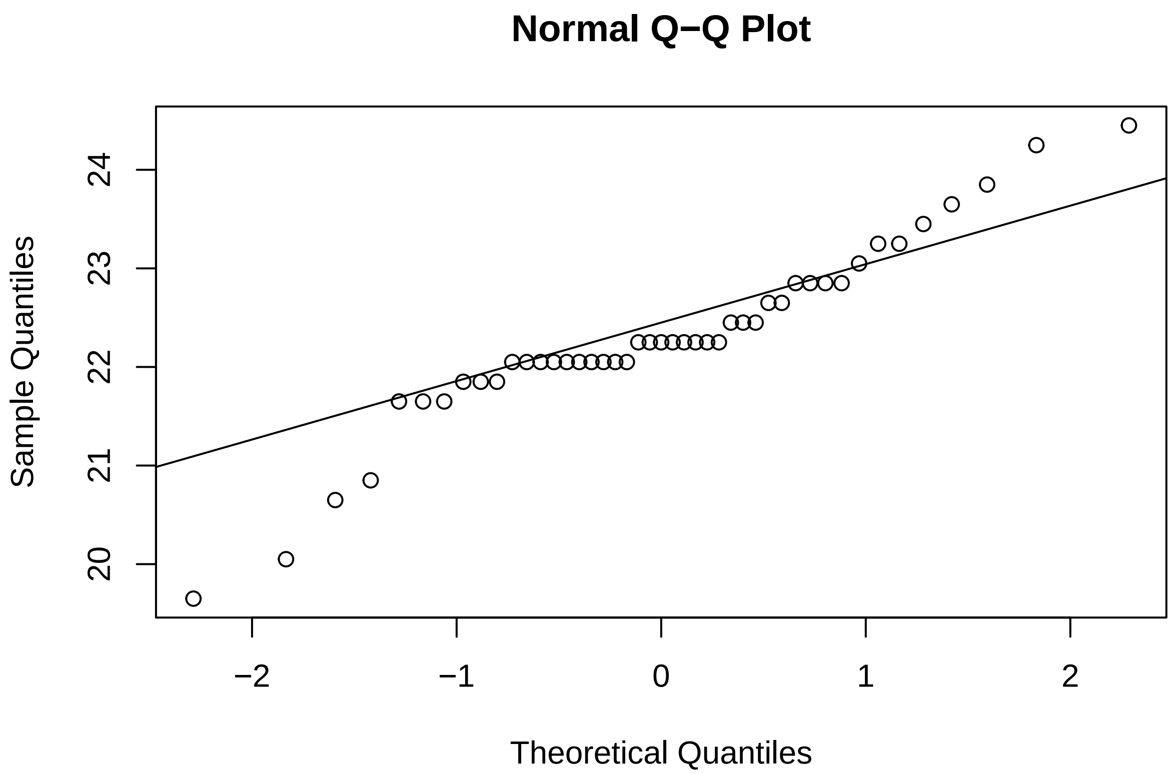
```
{qqnorm(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Hedge Sparrow"))
qqline(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Hedge Sparrow"))}
```



```
shapiro.test(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Meadow Pipit"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$Egg.Length, cuckoos$Host.Species == "Meadow Pipit")
## W = 0.93006, p-value = 0.009424
```

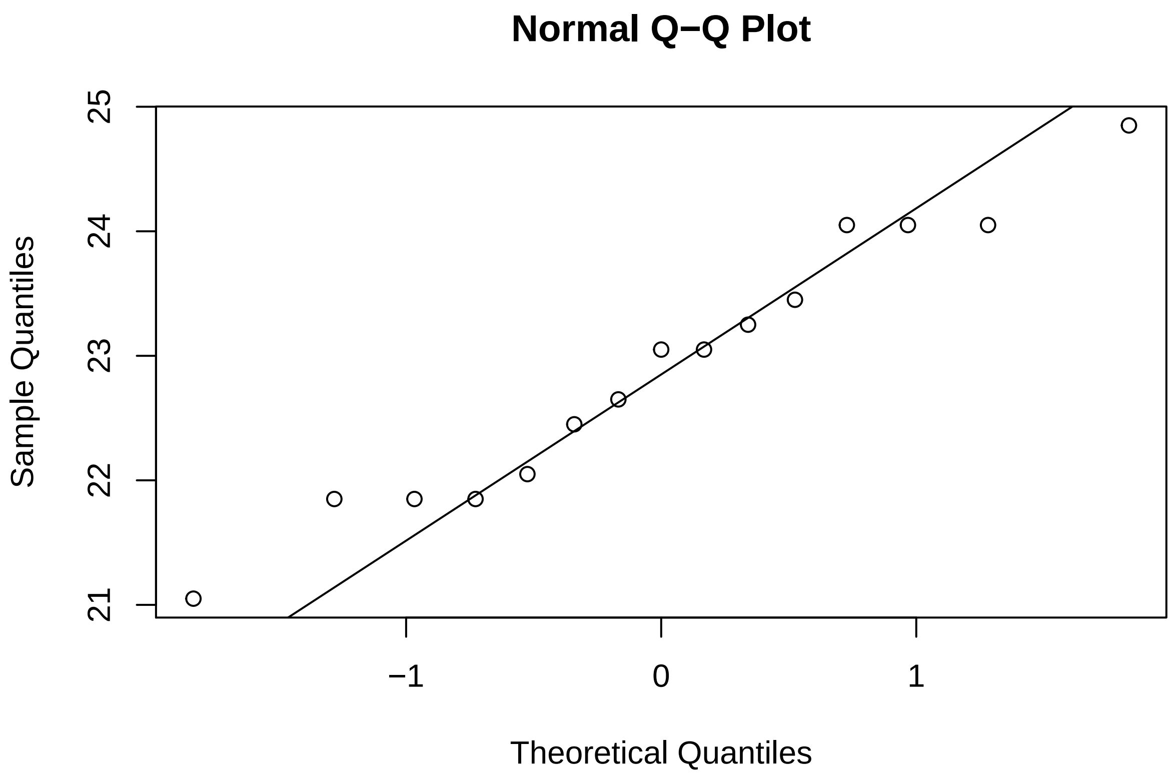
```
{qqnorm(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Meadow Pipit"))
qqline(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Meadow Pipit"))}
```



```
shapiro.test(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Pied Wagtail"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$Egg.Length, cuckoos$Host.Species == "Pied Wagtail")
## W = 0.96471, p-value = 0.7736
```

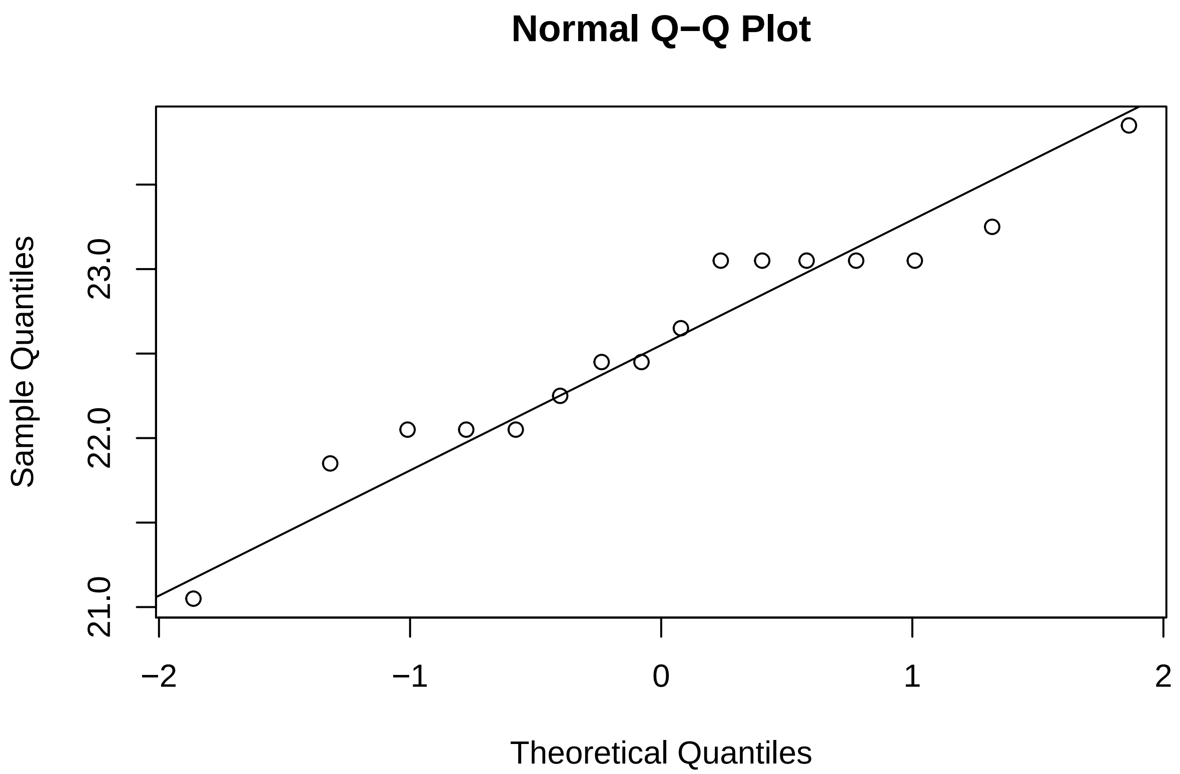
```
{qqnorm(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Pied Wagtail"))
qqline(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Pied Wagtail"))}
```



```
shapiro.test(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Robin"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$Egg.Length, cuckoos$Host.Species == "Robin")
## W = 0.95212, p-value = 0.5239
```

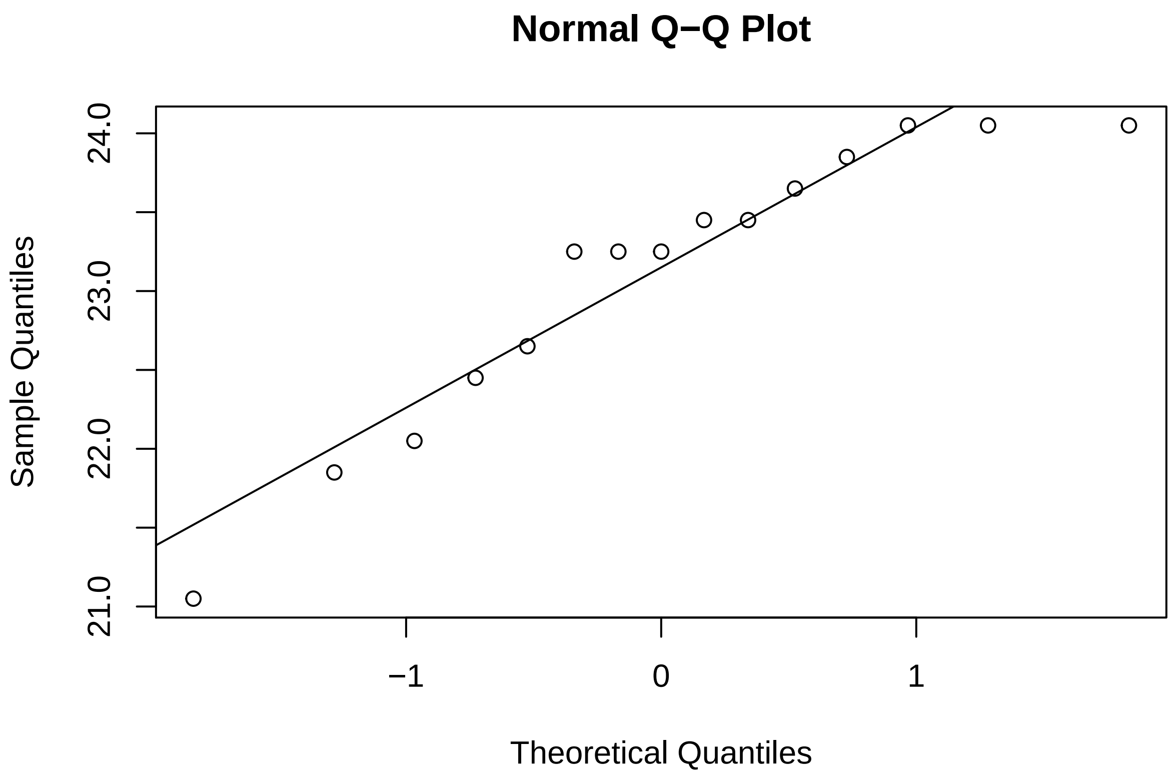
```
{qqnorm(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Robin"))
qqline(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Robin"))}
```



```
shapiro.test(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Tree Pipit"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$Egg.Length, cuckoos$Host.Species == "Tree Pipit")
## W = 0.89772, p-value = 0.08786
```

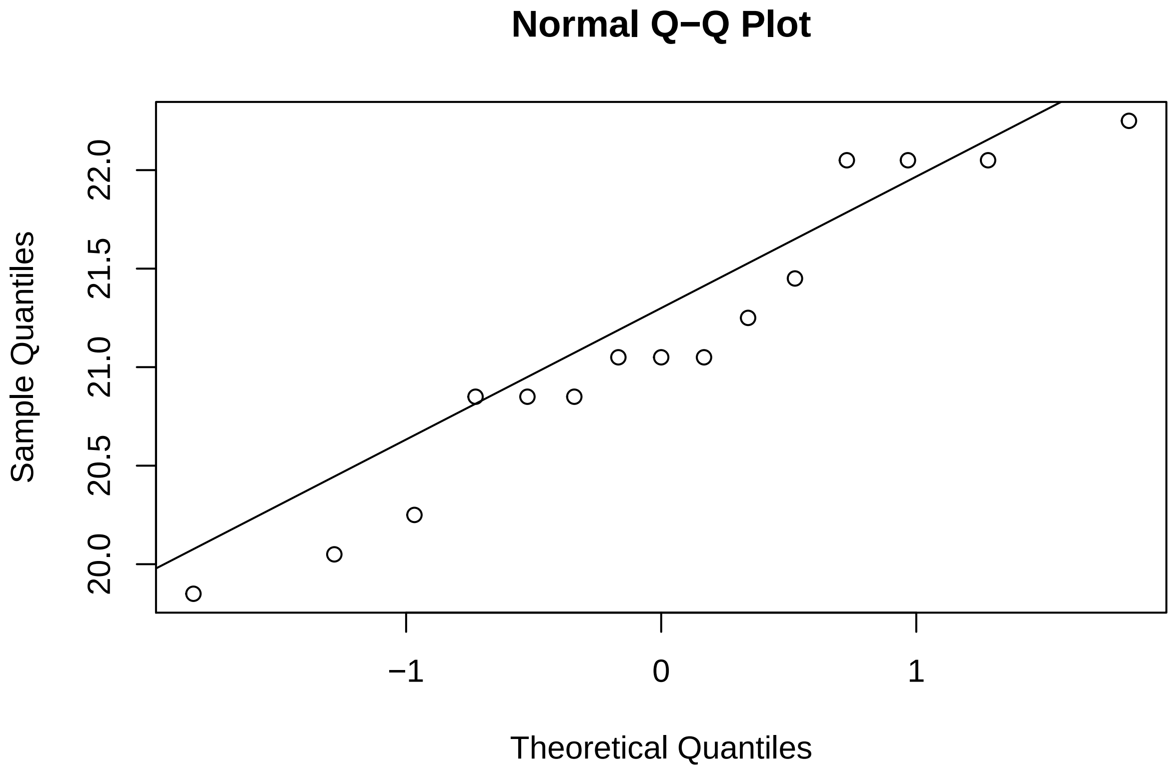
```
{qqnorm(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Tree Pipit"))
qqline(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Tree Pipit"))}
```



```
shapiro.test(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Wren"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$Egg.Length, cuckoos$Host.Species == "Wren")
## W = 0.93295, p-value = 0.3019
```

```
{qqnorm(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Wren"))
qqline(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Wren"))}
```



The null hypothesis of the Shapiro Wilk Normality test is that the population is normally distributed: any significant p-value allows us to reject the null hypothesis and conclude that the population is not normally distributed. In a Q-Q plot, data points form a straight line when the data is normally distributed.

Hedge-Sparrow-> Shapiro-Wilk: p=0.54, cannot reject null so the data is normally distributed, Plot: data points roughly form a straight line so the data is normally distributed.

Meadow Pipit-> Shapiro-Wilk: p=0.009, reject the null so data is not normally distributed, Plot: the data appears to have a nonlinear shape, so the data is not normally distributed.

Pied Wagtail-> Shapiro-Wilk: p=0.77, cannot reject null so the data is normally distributed, Plot: this plot appears more difficult to interpret, but based on the p-value provided by the Shapiro-Wilk, I would conclude that this shape is sufficiently linear to claim the data is normally distributed.

Robin> Shapiro-Wilk: p=0.52, cannot reject null so the data is normally distributed, Plot: this plot appears more difficult to interpret, but based on the p-value provided by the Shapiro-Wilk, I would conclude that this shape is sufficiently linear to claim the data is normally distributed.

Tree Pipit-> Shapiro-Wilk: p=0.09, cannot reject null so the data is normally distributed, Plot: this plot appears more difficult to interpret, but based on the p-value provided by the Shapiro-Wilk, I would conclude that this shape is sufficiently linear to claim the data is normally distributed.

Wren-> Shapiro-Wilk: p=0.30, cannot reject null so the data is normally distributed, Plot: this plot appears more difficult to interpret, but based on the p-value provided by the Shapiro-Wilk, I would conclude that this shape is sufficiently linear to claim the data is normally distributed.

Running the ANOVA test

Recall, the purpose of this research was to test whether the eggs laid by cuckoos were larger or smaller depending on which host bird's nest was being parasitized. The theory is, if cuckoos are able to mimic the host bird's own eggs, the host bird is less likely to notice the brood parasitism.

Question 4 Give the null and alternative hypotheses for an ANOVA which tests whether egg length differs between host species.

Null: The egg length is the same for all the host species. Alternative: The egg length is not the same for all the host species.

Question 5 Run an ANOVA using the function given in lecture slides, and save as a variable. Report the test statistic and P-value.

```
summary(aov(Egg.Length~Host.Species, data = cuckoos))
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## Host.Species    5  42.94   8.588   10.39 3.15e-08 ***
## Residuals     114  94.25   0.827                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cuckooANOVA <- aov(Egg.Length~Host.Species, data = cuckoos)
```

The test statistic is 10.39 with a p-value of 3.15e-08.

Question 6 Interpret the results of your ANOVA. In your answer, make sure to re-visit the initial question/experimental background, report the test statistic, and explicitly state what the p-value is the probability of.

Because $p < 0.05$, we reject the null hypothesis that all of the egg lengths are the same. The p-value represents the probability of getting an F statistic as big as 10.39 when the null hypothesis is true.

Question 7 Besides the assumption of independence and random sampling, what are the assumptions of the ANOVA test? Are they met? You will likely need to do additional coding to evaluate all the assumptions. Tip: if you are missing any necessary packages, install them using *install.packages*(”*packagename*”) in the console.

```
#1.The observations must be independent of each other and randomly sampled
#2. The residuals must be normally distributed (evaluate whether cuckoo egg length data is normally dis
#3. The groups must have roughly equal standard deviations (ratio of largest to smallest SD < 2)
#SD of Hedge-Sparrow
HS_sd <- sd(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Tree Pipit"), na.rm = FALSE)
#SD of Meadow Pipit
MP_sd <- sd(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Meadow Pipit"), na.rm = FALSE)
#SD of Pied Wagtail
PW_sd <- sd(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Pied Wagtail"), na.rm = FALSE)
#SD of Robin
R_sd <- sd(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Robin"), na.rm = FALSE)
#SD of Tree Pipit
TP_sd <- sd(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Tree Pipit"), na.rm = FALSE)
#SD of Wren
W_sd <- sd(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Wren"), na.rm = FALSE)
#Ratio between largest and smallest SD must be < 2.
ratio <- 1.07/0.68
```

The residuals must be normally distributed: Each of the residuals are normally distributed except for that of the Meadow Pipit, so this assumption is not fully met. The groups must have roughly equal standard deviations: all of the standard deviations are roughly the same, with the ratio of Robin (smallest, sd=0.68) standard deviation to Pied Wagtail (largest, sd=1.07) is 1.6 (less than the threshold of 2). This assumption is met.

Plotting the data

Graphical representation of the data will help you interpret the results of an ANOVA. The results of a statistical test should *always* be accompanied by an informative plot. In fact, usually making the plot is the first step you would do, before even running the test.

Consult the below link describing how to create stripcharts in R:

https://static1.squarespace.com/static/5eb33c095018927ea433a883/t/5f7a84444450c5069e7bba9e/1601864778198/Plotting-in-R.pdf

Question 8 Create a stripchart to display the egg length data from the cuckoo dataset. Remember to label your axes appropriately. Add error bars (sd or se) and points for each species' mean as demonstrated above. You will have to modify example code to account for 6 groups (species) whereas the examples shown above have different numbers of groups.

```
# Find each mean
mean.length <- aggregate(Egg.Length~Host.Species, cuckoos, FUN=mean)
head(mean.length)
```

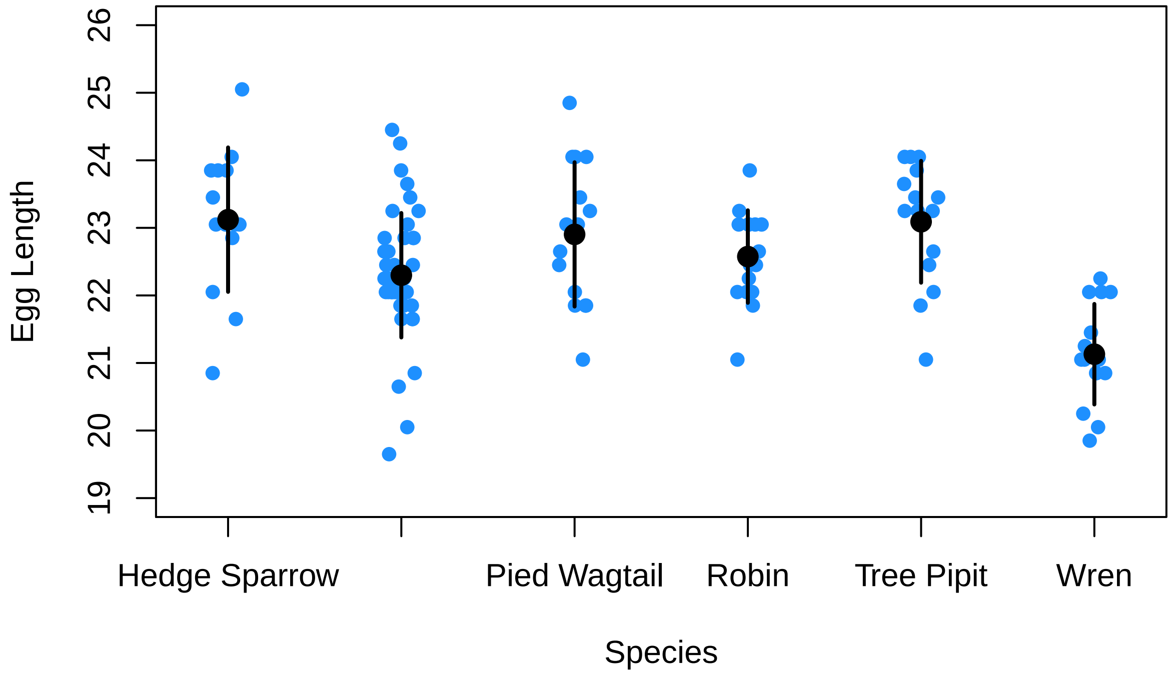
```
##    Host.Species Egg.Length
## 1 Hedge Sparrow   23.12143
## 2  Meadow Pipit   22.29889
## 3  Pied Wagtail   22.90333
## 4         Robin   22.57500
## 5    Tree Pipit   23.09000
## 6          Wren   21.13000
```

```
# Find each sd in a table
sd.length <- aggregate(Egg.Length~Host.Species, cuckoos, FUN=sd)

#Plot
{stripchart(Egg.Length~Host.Species, vertical= TRUE, method= "jitter", pch=16, col="dodgerblue", data=cu

#Plot means
points(mean.length$Egg.Length ~ c(1:length(levels(mean.length$Host.Species))), pch=16, col="black", cex=

#Plot SD
segments(x0 = c(1:length(levels(mean.length$Host.Species))), x1 = c(1:length(levels(mean.length$Host.Spe
}
```



Post Hoc Tests

Remember that an ANOVA tells you only whether the means differ among groups, not which exact groups differ from one another. To know which specific groups differ, we need to do post-hoc tests, which compare means among all pairs, accounting for multiple testing.

The function `TukeyHSD` does pairwise post-hoc tests to compare each pair of species. The basic code is: `TukeyHSD(yourmodelname)`. The p adj column gives a corrected P-value for that particular comparison.

Question 9 Run a Tukey test on the cuckoo ANOVA model (i.e. the variable you saved above). Are eggs laid in Tree-Pipit nests significantly different than those laid in Robin's nests?

```
TukeyHSD(cuckooANOVA)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Egg.Length ~ Host.Species, data = cuckoos)
##
## $Host.Species
##                                   diff          lwr         upr     p adj
## Meadow Pipit-Hedge Sparrow -0.82253968 -1.629133605 -0.01594576 0.0428621
## Pied Wagtail-Hedge Sparrow -0.21809524 -1.197559436  0.76136896 0.9872190
## Robin-Hedge Sparrow        -0.54642857 -1.511003196  0.41814605 0.5726153
## Tree Pipit-Hedge Sparrow   -0.03142857 -1.010892769  0.94803563 0.9999990
## Wren-Hedge Sparrow         -1.99142857 -2.970892769 -1.01196437 0.0000006
## Pied Wagtail-Meadow Pipit   0.60444444 -0.181375330  1.39026422 0.2324603
## Robin-Meadow Pipit          0.27611111 -0.491069969  1.04329219 0.9021876
## Tree Pipit-Meadow Pipit     0.79111111  0.005291337  1.57693089 0.0474619
## Wren-Meadow Pipit          -1.16888889 -1.954708663 -0.38306911 0.0004861
## Robin-Pied Wagtail         -0.32833333 -1.275604766  0.61893810 0.9155004
## Tree Pipit-Pied Wagtail     0.18666667 -0.775762072  1.14909541 0.9932186
## Wren-Pied Wagtail          -1.77333333 -2.735762072 -0.81090459 0.0000070
## Tree Pipit-Robin            0.51500000 -0.432271433  1.46227143 0.6159630
## Wren-Robin                 -1.44500000 -2.392271433 -0.49772857 0.0003183
## Wren-Tree Pipit            -1.96000000 -2.922428738 -0.99757126 0.0000006
```

No, eggs laid in Tree Pipet nests are not significantly different than those laid in Robin nests (p=0.6159630 > 0.05).

Question 10 Describe how an F statistic is calculated. What does a large F statistic indicate about your data vs. a small one?

An F statistic is a ratio of mean squares, an estimate for variance in a population, where F = variation between sample means / variation within sample means. These variances are calculated by squaring the standard deviation. A large F statistic indicates that the variability of group means is large relative to within group variation. A small F statistic indicates that the variability of group means is small relative to within group variation.

Next steps in research

BONUS (optional) Re-read the original background presented at the beginning of this lab. What might be an experiment you would run, or additional analyses you would want to conduct to further investigate whether variation in cuckoo egg size is actually an adaptation to disguise their eggs from the host?

I would be interested in running an experiment in the field to not only collect more data points, but also to test our hypothesis with more species, particularly those with either a comparatively large or small egg size to the rest of the species we have already collected date on.