Lab 7: Cuckoos

Jackson Anderson

Oct 22, 2020

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?

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse 1
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_conflic
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# your code here
cuckoos <- read_csv("~/Repos/School/BioStats/Lab7/cuckooeggs.csv")
```

```
## Parsed with column specification:
## cols(
##   `Host Species` = col_character(),
##   `Egg Length` = col_double()
## )
```

```
str(cuckoos)
```

```
## tibble [120 x 2] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Host Species: chr [1:120] "Hedge Sparrow" "Hedge Sparrow" "Hedge Sparrow" "Hedge Sparrow" ...
##  $ Egg Length  : num [1:120] 20.9 21.6 22.1 22.9 23.1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Host Species` = col_character(),
##   ..   `Egg Length` = col_double()
##   .. )
```

```
# code to convert the host-species variable to a "factor", i.e. categorical data
cuckoos$HostSpecies = as.factor(cuckoos$'Host Species')
cuckoos$EggLength <- cuckoos$'Egg Length'
cuckoos <- cuckoos[,-c(1,2)]
```

The predictor variable is 'Host Species' and the response is '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)
```

```
## tibble [120 x 2] (S3: tbl_df/tbl/data.frame)
##  $ HostSpecies: Factor w/ 6 levels "Hedge Sparrow",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ EggLength  : num [1:120] 20.9 21.6 22.1 22.9 23.1 ...
```

Six different host species of bird were 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)
#
# # Or
#
# shapiro.test(subset(cuckoos$Egg.Length, cuckoos$Host.Species=="Hedge Sparrow"))
```

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$EggLength, cuckoos$HostSpecies=="Hedge Sparrow"))
# qqline(subset(cuckoos$EggLength, cuckoos$HostSpecies=="Hedge Sparrow"))}
```

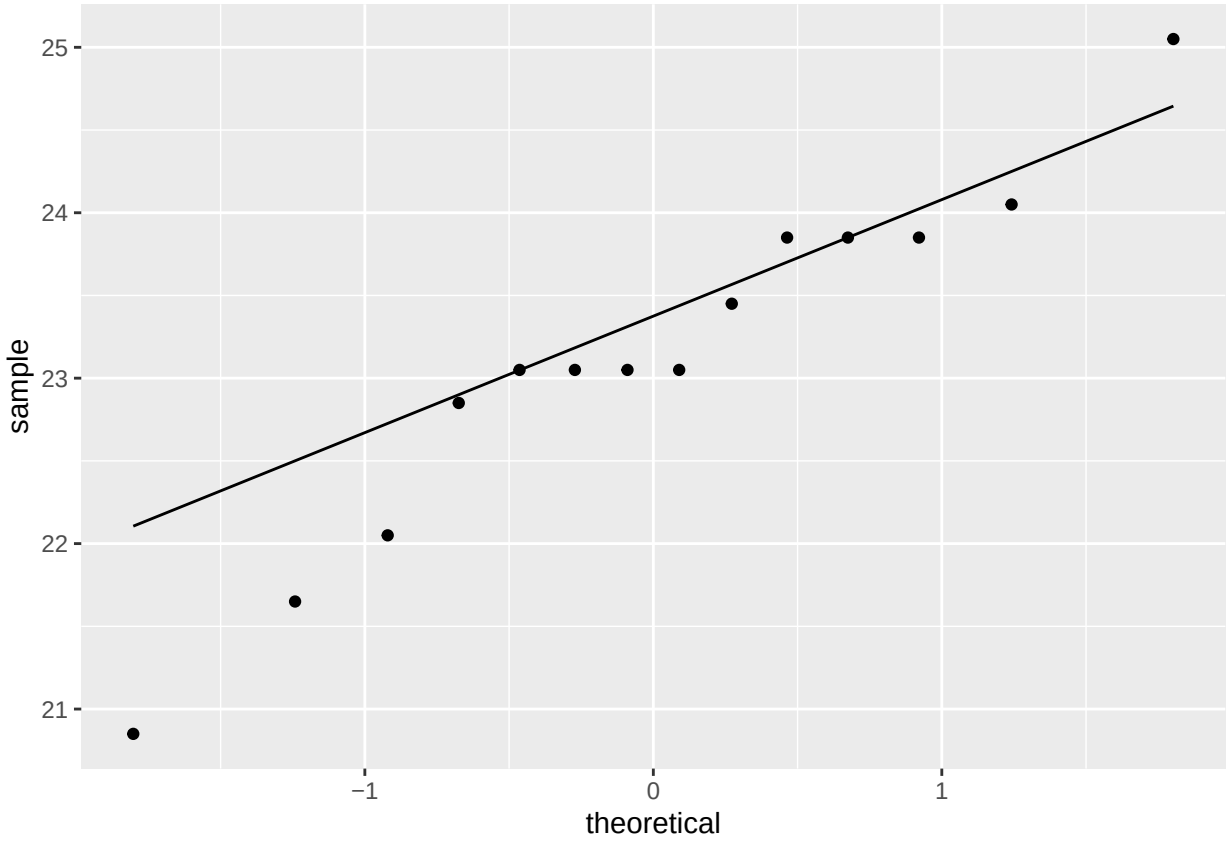
Question 3 Using the Shapiro-Wilk test for normality, as well as visual inspection of the data (i.e., plotting), evaluate whether cuckoo egg length data is normally distributed *within each group*. Interpret the output of the test.

```r
#response variable normality testing

#Hedge Sparrow
shapiro.test(subset(cuckoos$EggLength, cuckoos$HostSpecies=="Hedge Sparrow"))#Normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$EggLength, cuckoos$HostSpecies == "Hedge Sparrow")
## W = 0.94843, p-value = 0.5366
```

```r
hedgeDat <- subset(cuckoos, HostSpecies == "Hedge Sparrow")
ggplot(hedgeDat, aes(sample = EggLength)) +
  geom_qq() +
  geom_qq_line()
```

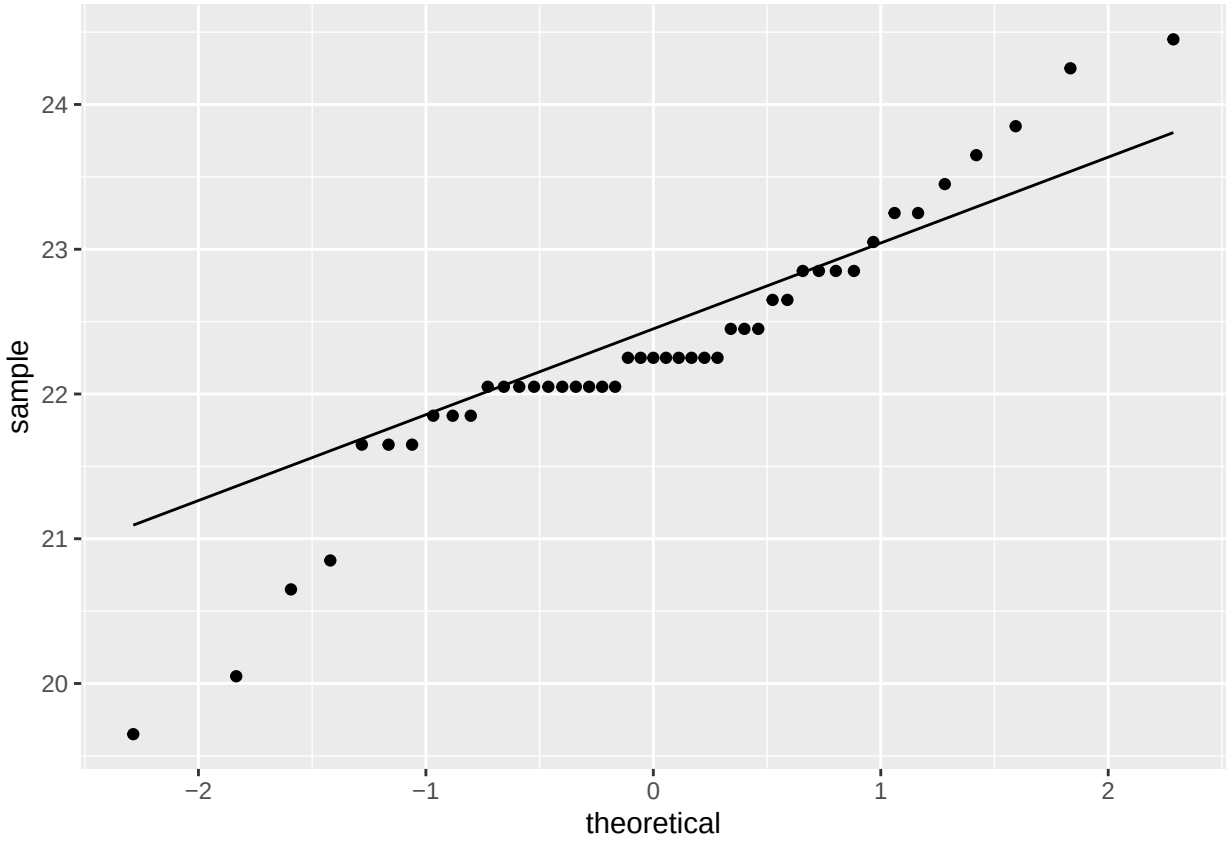


Normal

```r
#Meadow Pipit
shapiro.test(subset(cuckoos$EggLength, cuckoos$HostSpecies=="Meadow Pipit"))#Heteroskedastic
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$EggLength, cuckoos$HostSpecies == "Meadow Pipit")
## W = 0.93006, p-value = 0.009424
```

```
meadowDat <- subset(cuckoos, HostSpecies == "Meadow Pipit")
ggplot(meadowDat, aes(sample = EggLength)) +
  geom_qq() +
  geom_qq_line()
```

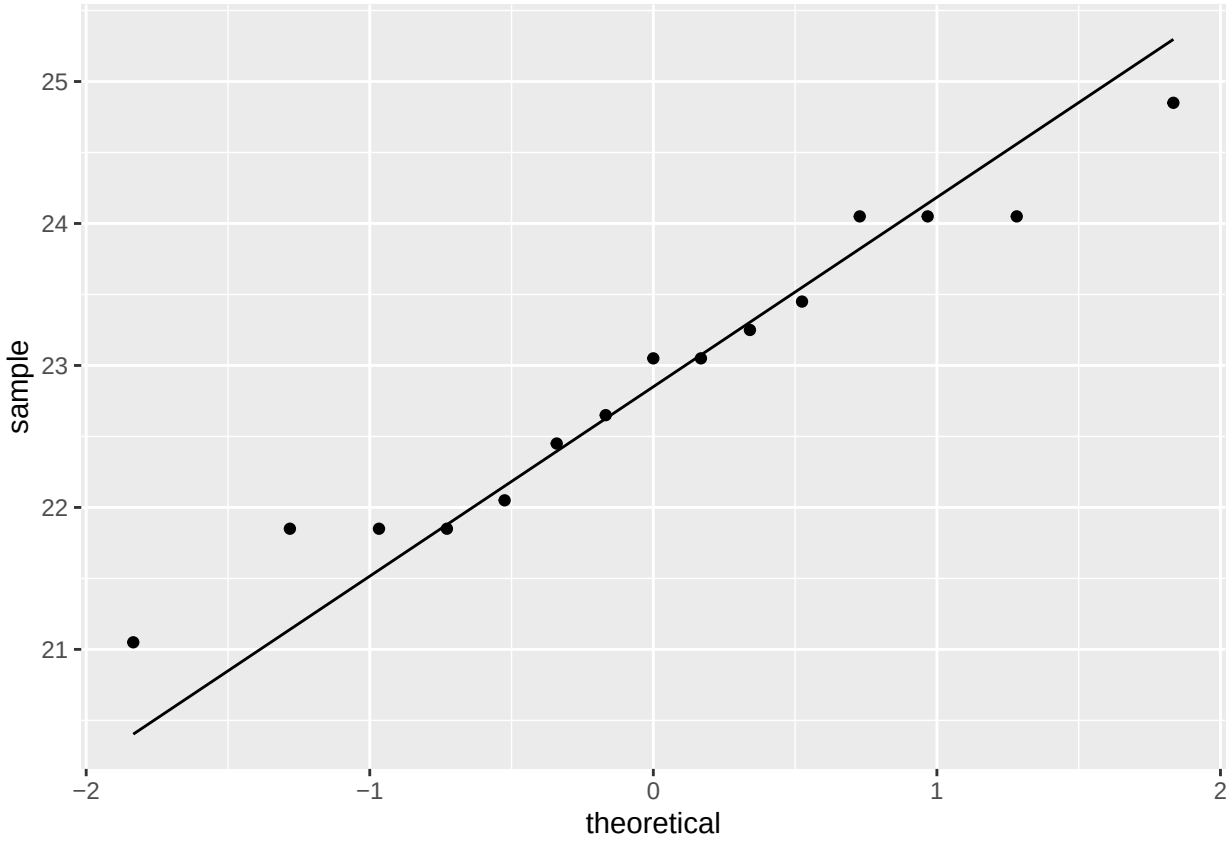


Heteroskedastic

```
#Pied Wagtail
shapiro.test(subset(cuckoos$EggLength, cuckoos$HostSpecies=="Pied Wagtail"))#Normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$EggLength, cuckoos$HostSpecies == "Pied Wagtail")
## W = 0.96471, p-value = 0.7736
```

```
piedDat <- subset(cuckoos, HostSpecies == "Pied Wagtail")
ggplot(piedDat, aes(sample = EggLength)) +
  geom_qq() +
  geom_qq_line()
```

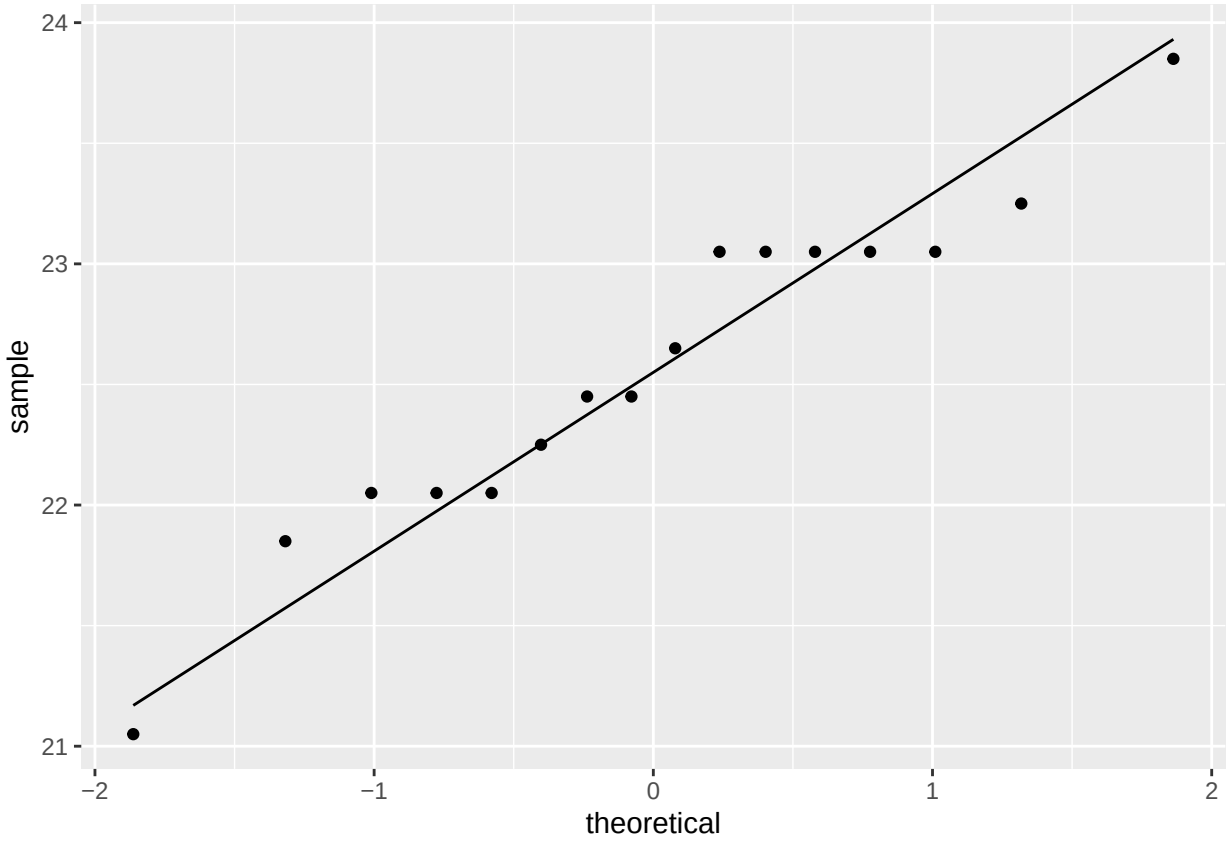


Normal

```
#Robin
shapiro.test(subset(cuckoos$EggLength, cuckoos$HostSpecies=="Robin"))#Normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$EggLength, cuckoos$HostSpecies == "Robin")
## W = 0.95212, p-value = 0.5239
```

```
robinDat <- subset(cuckoos, HostSpecies == "Robin")
ggplot(robinDat, aes(sample = EggLength)) +
  geom_qq() +
  geom_qq_line()
```

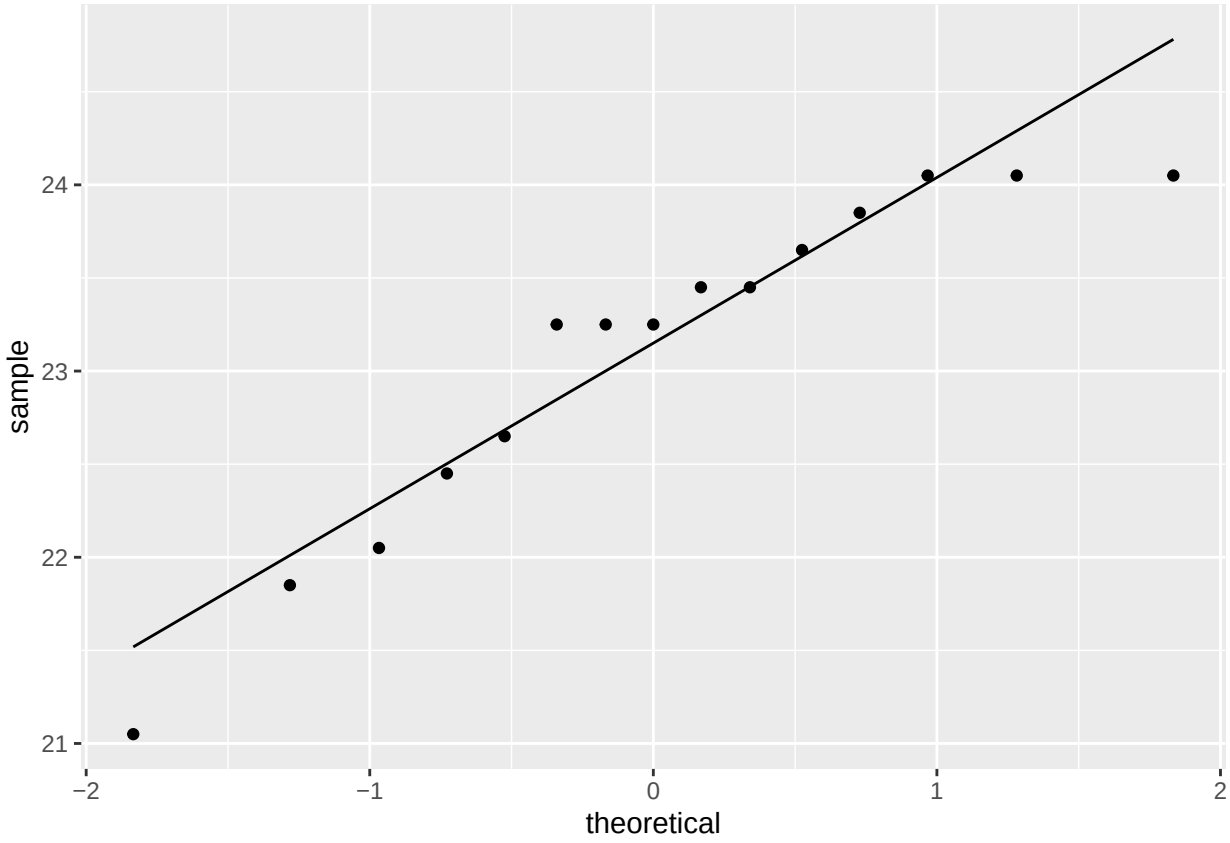


Normal

```
#Tree Pipit
shapiro.test(subset(cuckoos$EggLength, cuckoos$HostSpecies=="Tree Pipit"))#Normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$EggLength, cuckoos$HostSpecies == "Tree Pipit")
## W = 0.89772, p-value = 0.08786
```

```
treePipDat <- subset(cuckoos, HostSpecies == "Tree Pipit")
ggplot(treePipDat, aes(sample = EggLength)) +
  geom_qq() +
  geom_qq_line()
```

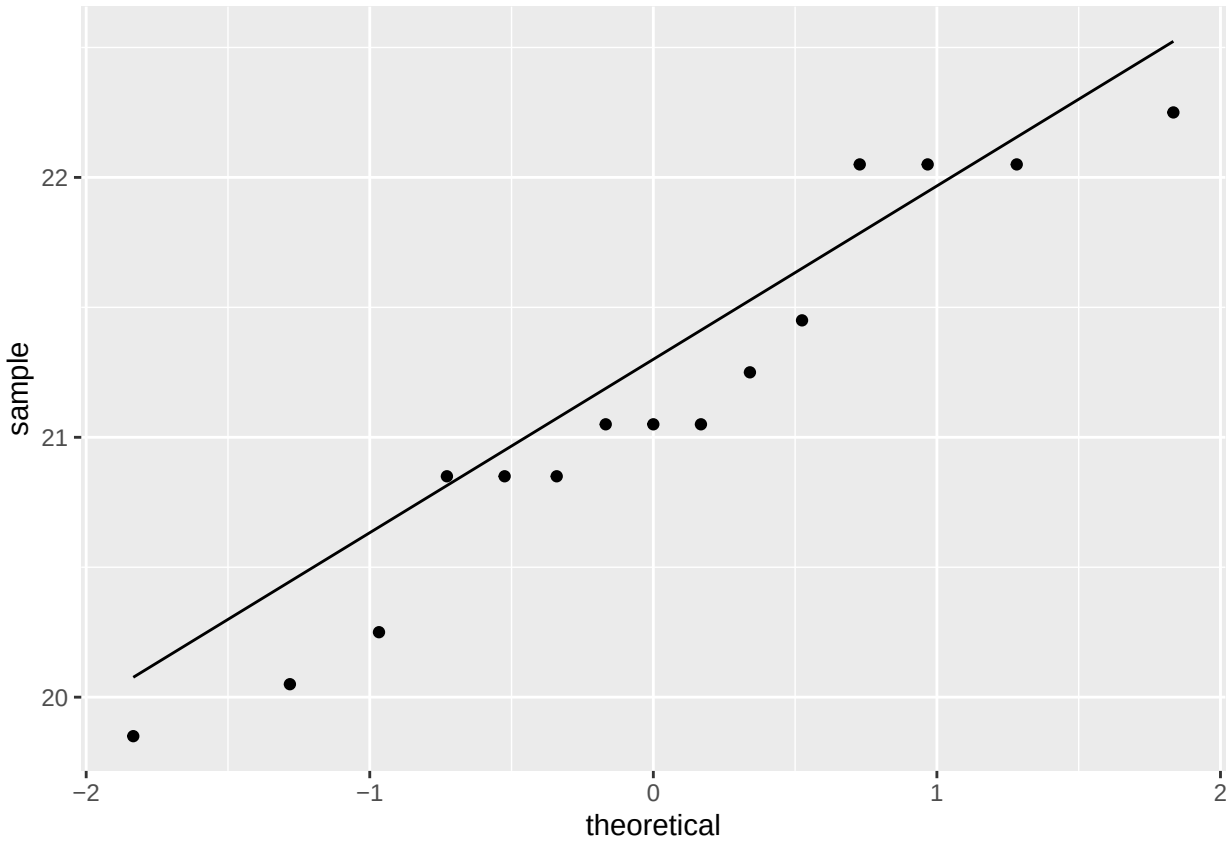


Normal

```
#Wren
shapiro.test(subset(cuckoos$EggLength, cuckoos$HostSpecies=="Wren"))#Normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(cuckoos$EggLength, cuckoos$HostSpecies == "Wren")
## W = 0.93295, p-value = 0.3019
```

```
wrenDat <- subset(cuckoos, HostSpecies == "Wren")
ggplot(wrenDat, aes(sample = EggLength)) +
  geom_qq() +
  geom_qq_line()
```



Normal

OVERALL: Egg length response variable is normally distributed within all treatments, aside from that of Meadow Pipit, where it was heteroskedastic

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: Host nest has no effect on Egg length Alternative: Egg length will differ significantly among host nest

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

```
birdAnova <- aov(EggLength ~ HostSpecies, data = cuckoos)
summary(birdAnova)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## HostSpecies   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
```

There was a significant biological effect; Host Species has a significant effect on Egg Length of parasatized eggs (p<.0001, F= 10.39

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.

The results of this test indicate that Host Species has a significant effect on the size of the cuckoo egg that is laid in its nest. These results indicate an overall treatment effect, but do not indicate where this difference is coming from, or which treatments contribute to this difference; pairwise comparisons would be needed to gather this information, which at this point seems necessary given the result of the experiment-wide anova. The test statistic (F stat) was 10.39, and our p-value is a probability of getting our observed signal to noise ratio (MSdifference among/MSdifference within groups) if the null hypothesis were true. In this case, it is very unlikely that we would observe such a ratio, and so it makes more sense to reject the null hypothesis

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.

```
#Assessing residual variance structures for each group

#1) checking response variable distribution:
shapiro.test(cuckoos$EggLength) #response variable is normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cuckoos$EggLength
## W = 0.98241, p-value = 0.1193
```

```
#2) bartlett's test for residual variance distribution
bartlett.test(EggLength ~ HostSpecies, data = cuckoos) #within-group variation is normally distributed
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  EggLength by HostSpecies
## Bartlett's K-squared = 4.4794, df = 5, p-value = 0.4826
```

Both response variable and residual variation is normally distributed

```
#Comparing group standard deviations:

#creating empty vector for each birdSpeces SD
hostSD <- rep(0, 6)
#Creating vector of bird species for funciton input below
birdSpecies <- levels(cuckoos$HostSpecies)
#Creating function to caclculate bird species standard deviation
sdFunc <- function(levs) {
  for(i in 1:length(levels(cuckoos$HostSpecies))) {
  hostSD[i] <- cuckoos %>%
  filter(HostSpecies == levs[i]) %>%
  summarise(sd(EggLength))
  }
  return(hostSD)
}
#Running the function
```

```r
sdFunc(birdSpecies)
```

```
## [[1]]
## [1] 1.068737
##
## [[2]]
## [1] 0.9206278
##
## [[3]]
## [1] 1.067619
##
## [[4]]
## [1] 0.6845923
##
## [[5]]
## [1] 0.9014274
##
## [[6]]
## [1] 0.7437357
```

#largest sd was that of species 1 (1.06); smallest was that of species 4 (.901). 1.06/.901 < 2; assumption for relatively equal group sd's met

Q7 Answer: The other assumpitons of an ANOVA are a) residuals are normaly distributed, and b) groups must have roughly equal standard deviations. The above coding indicates that the assumptions of (a) are met: response variable is normally distributed and the within group residual variances are normal, too. Likewise, (b) 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:

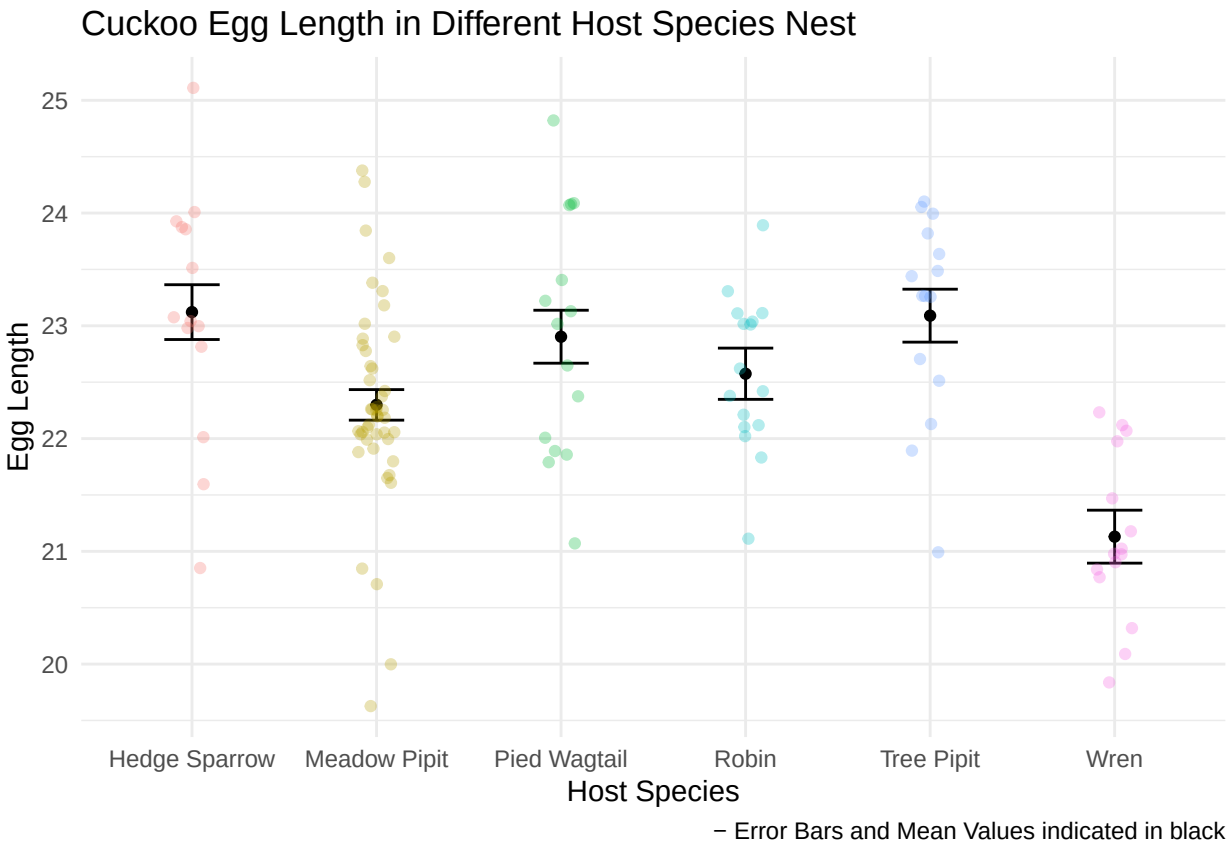
https://static1.squarespace.com/static/5eb33c095018927ea433a883/t/5f7a84444450c5069e7bba9e/1601864778198/Plotting-in-R.pdf

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.

```r
#creating datat frame with important summary stats
birdAov_noint <- lm(EggLength ~ 0 + HostSpecies, data = cuckoos)
summary_birdAov_noint <- summary(birdAov_noint)
importantStats <- as.data.frame(summary_birdAov_noint$coefficients)
importantStats$Species <- levels(cuckoos$HostSpecies)

ggplot(importantStats, aes(Species, Estimate)) +
  geom_point() +
  geom_errorbar(aes(ymin=Estimate -`Std. Error`, ymax=Estimate + `Std. Error`), width = 0.3, show.legend
  geom_jitter(data = cuckoos, aes(HostSpecies, EggLength, col = HostSpecies), width = .1, alpha = .3, sh
  theme_minimal() +
```

```
  labs(
    y = "Egg Length",
    x = "Host Species",
    title = "Cuckoo Egg Length in Different Host Species Nest",
    caption = " - Error Bars and Mean Values indicated in black"
  )
```



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(birdAnova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = EggLength ~ HostSpecies, data = cuckoos)
##
```

```
## $HostSpecies
##                                   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 the Tree Pipit nest are not sig. diff. from those laid in Robin nests

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 stat is calculated by dividing the calculated means sq. of among group variation by the calculated means sq. of within group variation. Given that an F-stat depends on the numDF and denDF, I suppose it is relative. However, a larger F-stat would indicate that there is much more among group variation relative to within group variation, and thus would suggest that the group means might not be equal. A smaller F stat would indicate a higher level of within group variation than among group variation, and so cases as such would suggest that the group means might not actually be different

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?

Your answer here