# Homework 13

Insert Name

Math 141, Week 13

**Due: 11:59pm, Friday April 29**

## Instructions

Work through the problems below and submit this document as a knitted .pdf to the Math 141 S22 Wells Lecture gradescope page.

**For each problem, put your solution between the bars of red stars.**

### Acknowledgements

If you work with a classmate, please write a note acknowledging this.

### Exercise 1

Data on a subset of 50 countries is loaded by the following code chunk

```
library(Lock5Data)
data("SampCountries")
```

Two of the variables in the dataset are life expectancy (`LifeExpectancy`) and percentage of government expenditure spent on health care (`Health`) for each country. We are interested in whether or not the percent spent on health care can be used to effectively predict life expectancy.

a. Create a scatterplot with regression line for `LifeExpectancy` as a function of `Health`.

b. Create a linear regression model and display the output using `get_regression_table`.

c. Use the regression table to write down a formula for the regression line. Interpret the slope in context.

d. Create a histogram of residuals as well as a residual plot to assess whether the conditions for making inference are met with these data.

e. Regardless of your answer, run a hypothesis test at the 0.01 level to determine whether the percentage of government expenditure on healthcare is a significant predictor of life expectancy.

f. Use `infer` to create a confidence interval for the correlation between `LifeExpectancy` and `Health`.

**Exercise 2**

Timber yield is approximately equal to the volume of a tree, however, this value is difficult to measure without first cutting the tree down. Instead, other variables, such as height and diameter, may be used to predict a tree's volume and yield. Researchers wanting to understand the relationship between these variables for black cherry trees collected data from 31 such trees in the Allegheny National Forest, Pennsylvania. Height is measured in feet, diameter in inches (at 54 inches above ground), and volume in cubic feet.

Data can be loaded with the following code chunk:

```
library(openintro)
data("cherry")
```

a. Create 3 scatterplots: volume vs height, volume vs diameter, and height vs diameter.

b. Compute the correlation between each pair of variables listed in part (a).

c. Fit 2 simple linear regression models, the first predicting volume as a function of height, and the second predicting volume as a function of diameter. Display the results using `get_regression_table`

d. In the models in part (c), is height a significant predictor of volume at the 0.01 level? Is diameter a significant predictor at the 0.01 level? Explain.

e. Create a multilinear model for volume as a function of both height and diameter. In the multilinear model, are both variables still significant at the 0.01 level? Explain why this doesn't contradict your answer to part (d).

f. Construct a residual vs predicted values plot, along with a histogram of residuals. Based on the plots, do you have any concerns about performing statistical inference for this data?

**Exercise 3**

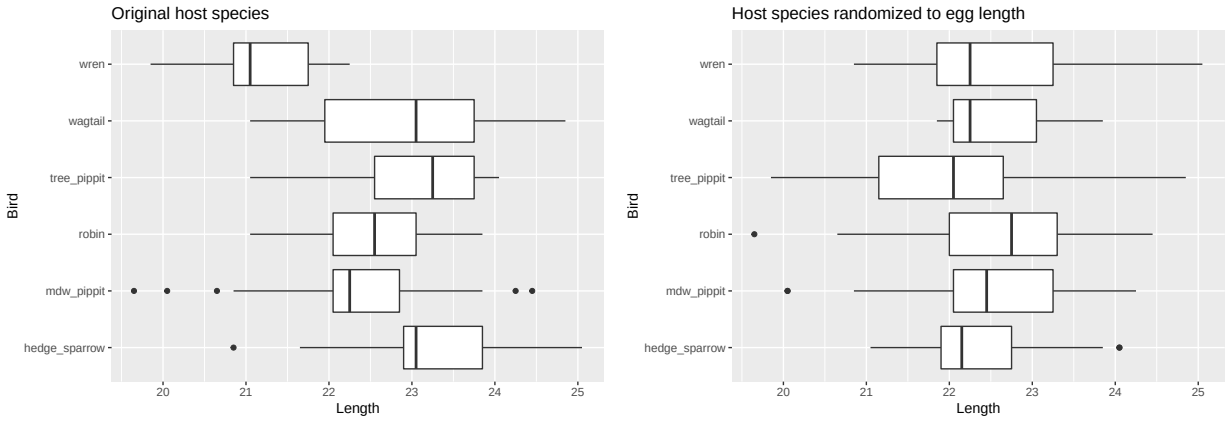
Cuckoo birds lay their eggs in other birds' nests, making them known as brood parasites. One question relates to whether the size of the cuckoo egg differs depending on the species of the host bird. Data for a random sample of 120 Cuckoo eggs can be loaded with the following code chunk.

```
cuckoo <- read_csv("https://reed-statistics.github.io/math141s22-wells-website/data/cuckoo.csv")
```

Consider the two boxplots shown below, one representing the original data, and the second representing data where the host species has been randomly assigned to the egg length:

```
set.seed(1001)
cuckoo %>% ggplot(aes(x = Length, y = Bird))+
  geom_boxplot()+
  labs(title = "Original host species")

cuckoo %>%
  sample_n( size = 120, replace = F) %>%
  mutate(Length = cuckoo$Length) %>%
  ggplot(aes(x = Length, y = Bird))+
  geom_boxplot()+
  labs(title = "Host species randomized to egg length")
```




a. Is the average egg length for the original data: more variable, less variable, or about the same as the randomized species? Justify your answer by appealing to specific features of the plots.

b. Consider the standard deviation of egg length within each species. Is the within species standard deviation of egg length for the original data: bigger, smaller, or about the same as the randomized species?

c. Recall that the F statistic's numerator measures how much the groups vary, while the denominator measures how much the within species values vary. Based on the plots, which data set likely has the larger F statistics, the original data or the randomized data? Explain.

d. Compute the mean and standard deviations of egg length within each bird species, along with the number of observations for each species.

e. Use `infer` to perform a hypothesis test to assess whether egg length and bird species are independent. Be sure to report your p-value and interpret your results in context of the data.