Kumar - SDS 291 Homework 2

HOMEWORK 2:

SET UP

```
library(Stat2Data)
library(tidyverse)
```

```
data("LeafWidth")
head(LeafWidth)
```

```
##       Width   Length    LWRatio     Area Year
## 1 0.6578947 85.13158 129.40000 70.32548 1974
## 2 0.9210526 53.15789  57.71429 48.96122 1918
## 3 0.9210526 59.86842  65.00000 57.72161 1993
## 4 0.9210526 63.68421  69.14286 37.36150 1946
## 5 0.9210526 64.21053  69.71429 62.55194 1956
## 6 0.9210526 86.97368  94.42857 67.52078 1964
```

EXERCISE 1

Use a t-test to support your answer to the research question: Is there a linear relationship between the width and length of a Hopbush plant’s leaves? In your analysis, be sure to treat the leaf width as the outcome variable.

```
leaf_model = lm(Width ~ Length, data = LeafWidth)
summary(leaf_model)
```

```
##
## Call:
## lm(formula = Width ~ Length, data = LeafWidth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8990 -1.0488 -0.3194  0.8751  5.6922
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.264938   0.446616   5.071  7.7e-07 ***
## Length      0.015176   0.007304   2.078   0.0388 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 1.457 on 250 degrees of freedom
## Multiple R-squared:  0.01697,    Adjusted R-squared:  0.01304
## F-statistic: 4.316 on 1 and 250 DF,  p-value: 0.03877
```

The p-value from this model tells us that there is less than a 5% chance that the **correlation** (slope) between these two variables occurred randomly if there was no linear relationship between them.

We can calculate a t-value using the following formula: (Observed Slope - Hypothesized Slope)/Standard Error of Slope. Under our null hypothesis that there's no relationship between leaf Length and Width, our Hypothesized Slope is 0. Thankfully we don't have to compute this manually, as the `summary()` function provides these values and calculates a t-statistic for us. Thus, we're told that our observed slope is 2.078 times more extreme than our expected slope under the null hypothesis (0). We're also told that the probability of that occurring randomly (the p-value) is 0.0388, which is less than 0.05 and thus provides sufficient evidence for us to reject the null hypothesis.

Therefore, based on this sample, we can conclude that there is evidence for a linear relationship between the length and width of Hopbush leaves.

EXERCISE 2

Exercise 2A: Compute a 99% confidence interval around the slope coefficient of the model you fit in Exercise 1. Compute manually: a confidence interval is computed by finding the marginal error and adding/subtracting it from the slope co-efficient, wherein:

$$ME = t^* * SE_X$$

```
## GIVENS
slope_coeff = 0.015176  ## from above
alpha = 0.01            ## because we're working with a 99% conf level
std_error = 0.007304

## CALCULATED VALUES
deg_free = nrow(LeafWidth) - 2

t_crit = qt(1 - alpha/2, deg_free)

margin_error = t_crit * std_error
```

Now that we have the margin of error, we can generate the lower and upper bounds of the interval

```
lower = slope_coeff - margin_error
upper = slope_coeff + margin_error

lower
```

```
## [1] -0.003782537
```

```
upper
```

```
## [1] 0.03413454
```

Manual calculation to check our work:

```
confint(leaf_model, level = 0.99)
```

```
##                    0.5 %     99.5 %
## (Intercept)  1.10568450 3.42419172
## Length      -0.00378389 0.03413573
```

Exercise 2B: Take this opportunity to explain to your colleagues what interpretation of confidence intervals the phrase "95% confident" supports. In this example, Tamara states: "We can be 95% confident the slope ranges from 15.7 to 22.1".

Kiera states: "Interesting, this means that there is a 95% chance the slope coefficient for the entire population of values is between 15.7 and 22.1"

Kiera is correct out of the two of them. The correct interpretation of "95% confident" in this case is as follows. When you construct a 95% confidence interval, you generate two values between which there is a 95% chance that, based on your sample, the actual slope (in this case) of the **population** exists between. Tamara's statement is too general, not specifying which slope or under what circumstances she's talking about.

EXERCISE 3

The original study came away with a t-value of 2.078, whereas this study has one of 1.066. Because the two studies have the same number of samples (and thus degrees of freedom, which inform t-value calculations), we can compare these numbers. t-values are a measure of how extreme the distance an observed slope is from our expected slope. With a null hypothesis that there's no relationship between the two variables, we expect the slope to be 0. These t-values indicate that the slopes are 2.078 and 1.066 times more extreme than what we expect under the null hypothesis. However, the sample of leaves from India has a t-value closer to 0, indicating a less extreme relationship between leaf length and width. This would be reflected with a larger p-value if one was generated.

EXERCISE 4

load data:

```
faithful %>%
  head()
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

Exercise 4A Use simple linear regression to obtain a "best guess" for this average wait time. Then, construct and interpret in context an appropriate interval (at 95% confidence) for that guess.

Load data

```r
faithful %>%
  head()
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

Generate a linear model

```r
faithful_model = lm(waiting ~ eruptions, data = faithful)
summary(faithful_model)
```

```
##
## Call:
## lm(formula = waiting ~ eruptions, data = faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0796  -4.4831   0.2122   3.9246  15.9719
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.4744     1.1549   28.98   <2e-16 ***
## eruptions    10.7296     0.3148   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.914 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

This model summary tells us the estimated increase in wait-time per minute of eruption is 10.7296 minutes. Based on this, we can construct a best guess for the average wait-time after a 4 minute eruption:

```r
estimate = 10.7296 * 4 + 33.4744
estimate
```

```
## [1] 76.3928
```

we can now generate a confidence interval for this estimate (based on a 4 minute eruption)

```r
predict(faithful_model, newdata = data.frame(eruptions = 4), interval = "predict",
        conf_level = 0.95)
```

```
##        fit      lwr     upr
## 1 76.39296 64.72382 88.0621
```

This interval tell us there's a 95% chance that the real waittime after a 4 minute eruption is between 64.72382 and 88.0621 minutes. Our estimate was 76.3928 minutes, which lies within our interval.

Exercise 4B This question is asking us to predict how long after a 4 minute eruption our friend would have to grab snacks, at a 95% confidence level. We can just modify the code used in Exercise 4A, so 95% of the time, after watching a 4 minute eruption, our friend would have between 75.62 and 77.17 minutes to get our snacks:

```
predict(faithful_model, newdata = data.frame(eruptions = 4), interval = "confidence",
        conf_level = 0.95)
```

```
##        fit     lwr      upr
## 1 76.39296 75.6189 77.16702
```

Exercise 4C These numbers are different, which makes sense because in one case, we are constructing a prediction interval (calculates the accuracy of a specific y-value predicted from the interval) in one and a confidence interval (estimates the accuracy of the mean response for any value X) in the other.