MATH 338

FINAL EXAM

SOFTWARE PORTION

DUE: WED, MAY 16, 2018 at 9:30 PM

## Exam Rules

You may refer to your textbook, any notes/code you wrote, and anything on Titanium. You may refer to other books, other Sapling resources, or the wider Internet, but if you use these resources to help you answer questions, you must cite them properly.

You may ask Dr. Wynne to clarify what a question is asking for, or to help you troubleshoot RGuroo errors and/or debug your R code. You may not ask any other people for help (on- or offline).

For full credit, include all R code (if using RStudio), graphs, and output. Save your answers as a .docx or .pdf file and upload the file to Titanium.

## Honor Statement

I certify that all work on this exam is my own, and that I have neither given nor received unauthorized help on the exam.

Name: ______________________

Date: ______________________

# Exam Instructions

This portion of the exam consists of three problems. Read the entire text of each problem and import the associated dataset into your software of choice. Then, for each problem:

A) [1 pt] Perform background analysis to determine the most appropriate inferential procedure to use to answer the question. Write the procedure you will use. Be as specific as possible.

B) [1.5 pts] Perform background and exploratory data analysis to determine whether the assumptions of the procedure you identified in part (A) have been met.

C) [0.5 pts] State explicitly whether all assumptions have been met or at least not terribly violated.

For **one** of the problems, there is a major violation of at least one assumption. **Do not perform any inference for this problem. Simply state that the suggested inferential procedure is not appropriate and explain why.**

For the remaining **two** problems:

D) [3 pts] Using a default 95% confidence level/5% significance level, perform the inferential procedure from part (A) in the statistical software, and answer the question.

You will receive 1 point for pasting relevant code and output (including error messages if you cannot get the code/dialog to work properly). The remaining 2 points will be earned for correctly restating the relevant part(s) of the output (for example, test statistic and p-value) and writing an appropriate interpretation/conclusion in the context of the question.

# Exam Problem 1

By convention, a study should have a statistical power of at least 80% (0.8) to be considered a "good study." However, it is not uncommon for studies to be published despite having much lower power.

The file rheumatoid_arthritis_power.csv contains the following variables for 34 studies of risk factors for rheumatoid arthritis:

- Marker: the genetic or lifestyle risk factor being studied
- Year: the year the meta-analysis reporting the study was published
- Author: the first author of the meta-analysis reporting the study
- Power: the estimated power of the study investigating the marker (reported as a proportion)

You can consider Year and Author to jointly describe the "source" of the information used to estimate the power for each study.

Assuming that this sample is representative of all studies published about risk factors for rheumatoid arthritis, do rheumatoid arthritis studies, on average, achieve 80% power?

**Important Note: This question does not ask you to compute power. Power is the variable of interest.**

A) What inferential procedure would you use to answer this question?

Best answer: one sample t-test for means, with $H_0$: $\mu = 0.8$ and $H_a$: $\mu < 0.8$

Also acceptable: one-sample t confidence interval for means

0.75 pts: one-sample t-test for means with an incorrect or unspecified null/alternative hypothesis pair

0.5 pts: anything less specific involving t

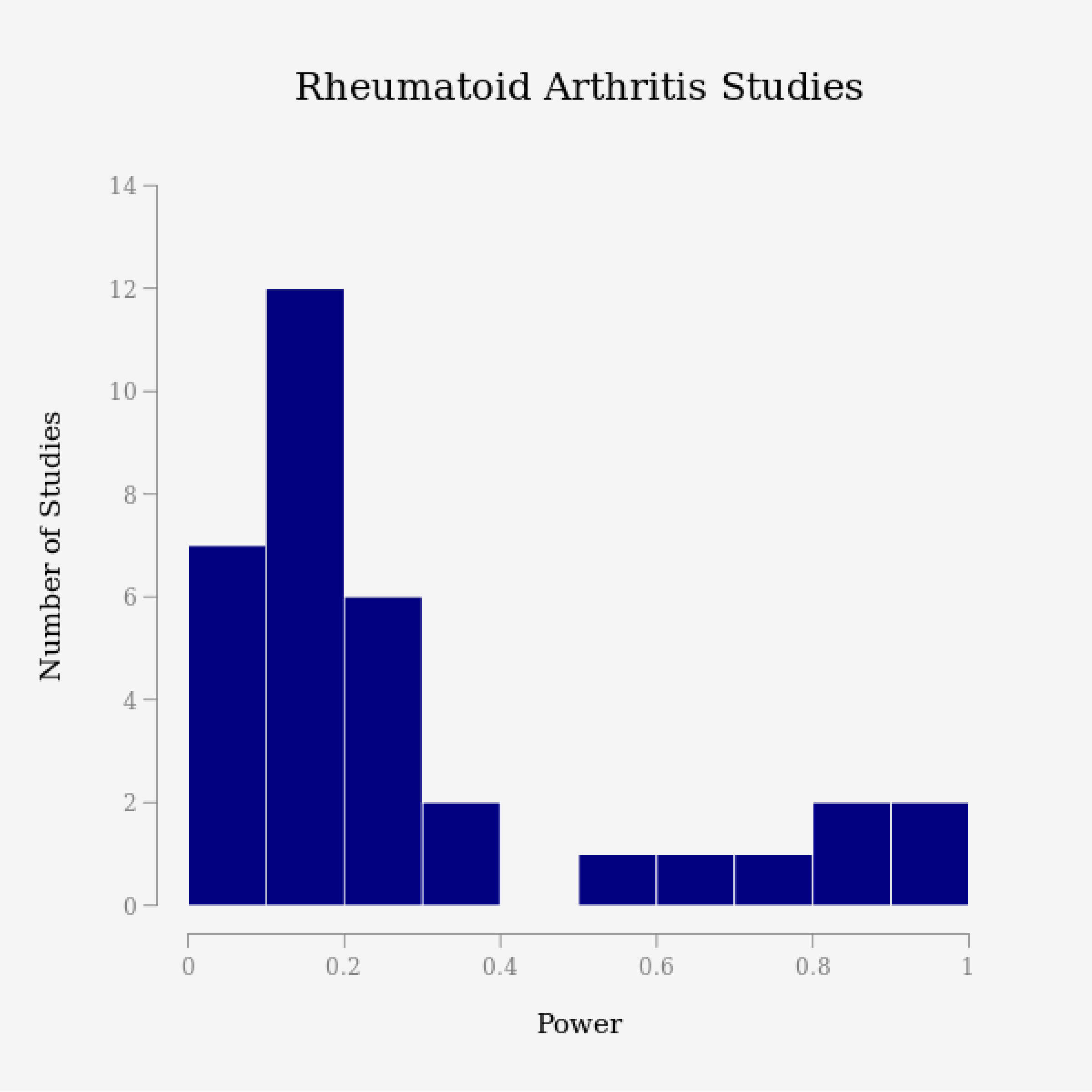
B) Perform background and exploratory data analysis to determine whether the assumptions of the procedure have been met. Write and paste your analysis below.

Check that a one-sample t test is appropriate:

Do we have a simple random sample? Based on the information given, it's not 100% clear that the studies are actually independent. If you view the actual data in R or Rguroo, there seem to be a lot of very similar markers that might have been studied using the same data.

Do we know the population standard deviation? No, but we have a dataset that we can use to estimate it.

Do we meet the sample size rule of thumb?

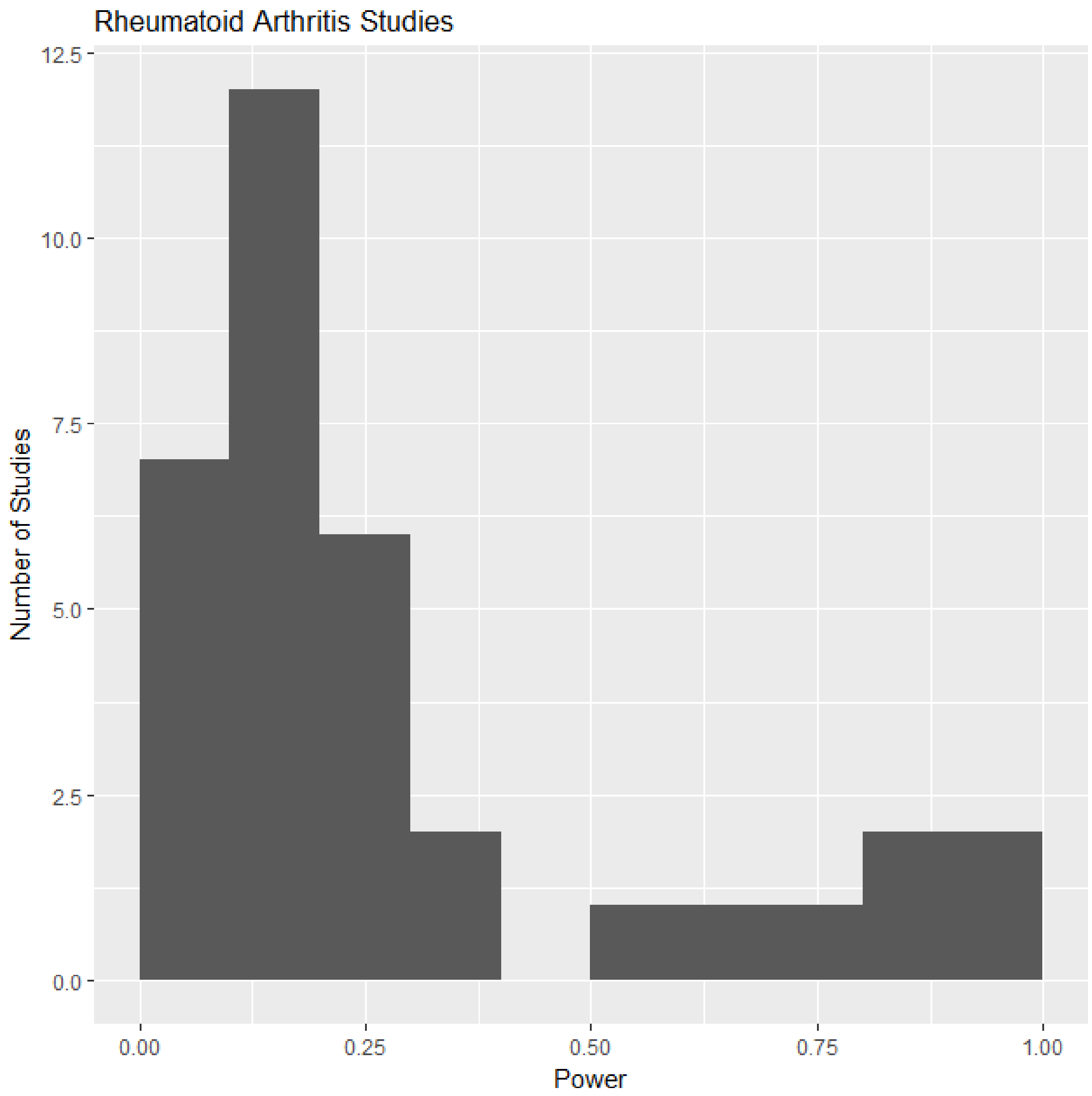


We have only 34 studies, and either there are six outliers or this distribution is heavily skewed right. Furthermore, three of those six "outliers" are for some kind of smoking risk factor, which is different from all of the genetic risk factors that have really low power.

Using R to generate the histogram:

```
> RA_power_histogram <- ggplot(rheumatoid_arthritis_power, aes(x =
Power)) + geom_histogram(center = 0.05, binwidth = 0.1) + labs(x =
"Power", y = "Number of Studies", title = "Rheumatoid Arthritis
Studies")

> print(RA_power_histogram)
```



C) Are all assumptions met (or at least not terribly violated)? If not, explain the violation(s).

Assumptions are not met. We are very leery of claiming the observations are independent, and even if we could, the sample size rule of thumb for a one-sample t test is not met.

D) If assumptions are violated, stop here. If assumptions are okay, perform the inference:

Assumptions are violated. We should not perform inference.

# Exam Problem 2

Some biologists are interested in the relationship between the length and width of an animal's mouth, because that relationship can influence the animal's diet. Although the length and width are obviously both dependent on the animal's size, it is reasonable to assume that the relationship does not change in mature animals of the same species.

The file crocs.csv contains the following variables, measured on 25 American crocodiles (*Crocodylus acutus*):

- Length: the length of the mouth, in mm
- Width: the width of the mouth, in mm

Assuming that this sample is representative of all American crocodiles, estimate the population mean mouth width of all American crocodiles whose mouths are 600 mm in length.

A) What inferential procedure would you use to answer this question?

Best answer: confidence interval for a mean response (in the simple linear regression framework)

Also acceptable: t confidence interval for mean (only if rest of analysis is clearly working in simple linear regression framework)

0.5 points: t confidence interval, t confidence interval for mean (if rest of analysis is unclear)

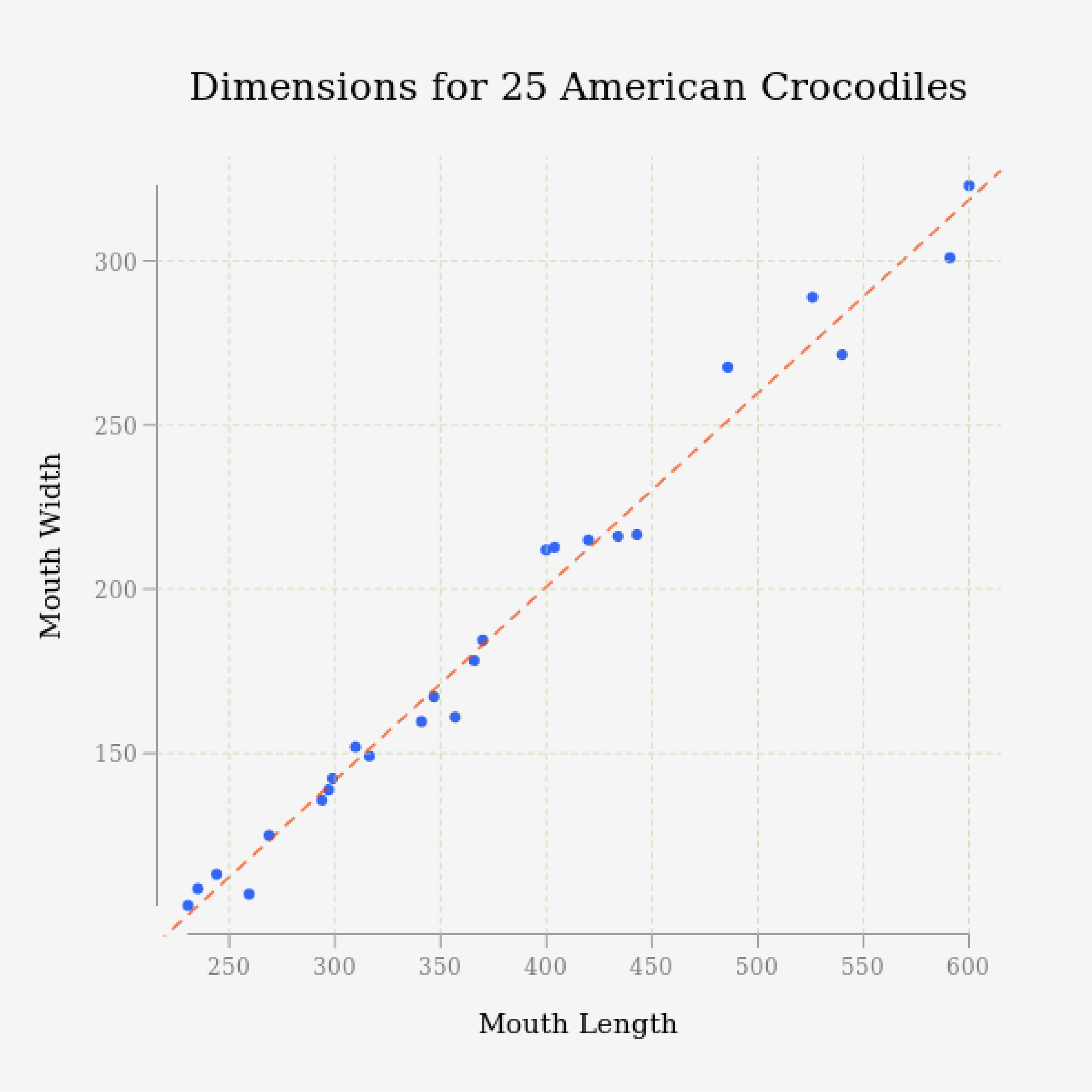
0.5 points: recognizing simple linear regression framework but not doing correct inference in it

B) Perform background and exploratory data analysis to determine whether the assumptions of the procedure have been met. Write and paste your analysis below.

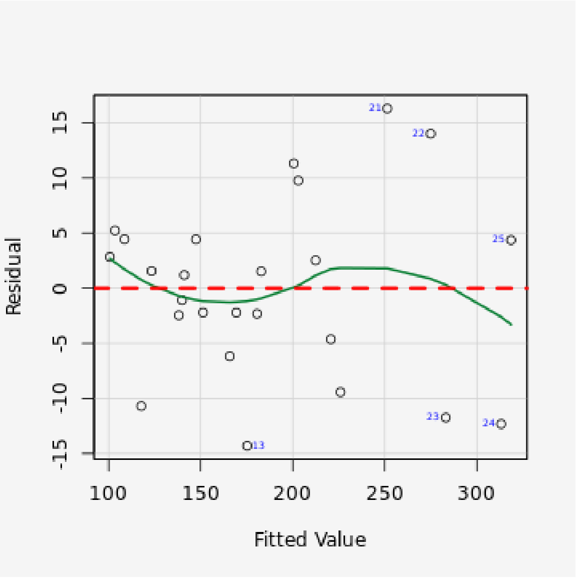
Check to see whether inference for simple linear regression is appropriate:

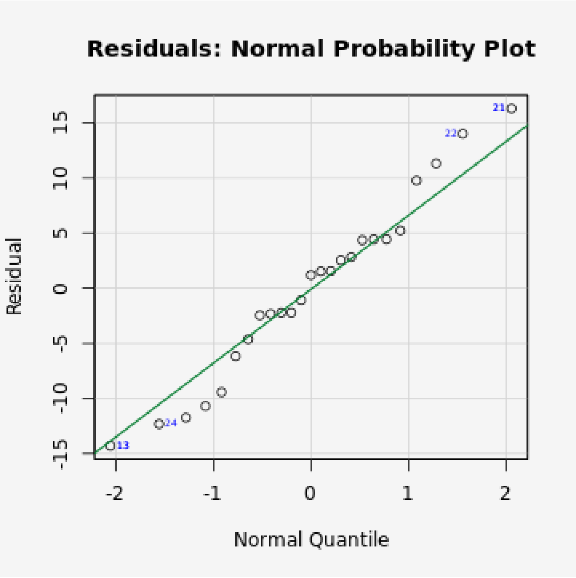
In the absence of any knowledge about how these crocodiles were sampled, it is probably reasonable to assume that residuals are independent.

We need to check our diagnostic plots to determine whether the linear model is appropriate, and whether the residuals are approximately N(0, σ).



(Least-square line not absolutely necessary for checking linearity but it helps)



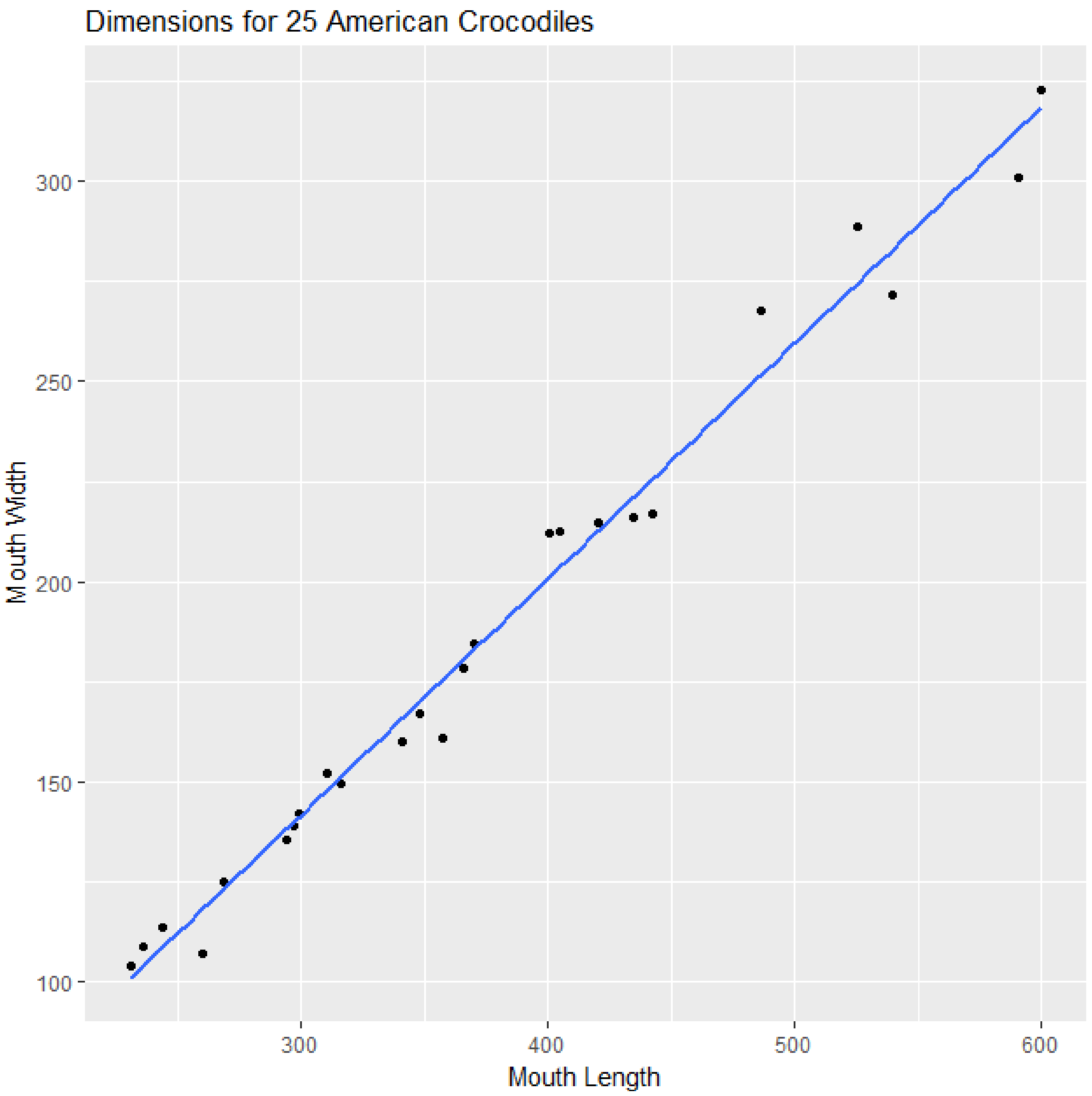


Linear model is appropriate. Residual plot shows no obvious trend. There might be a little bit of fanning but it's not terrible, and the q-q plot looks okay (not great, but okay). If anything, the problem seems to be with the choice of points to draw the line through.

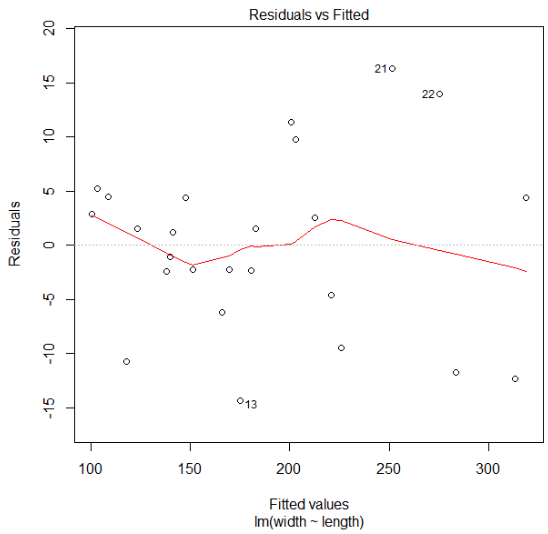
Using R to generate the plots:

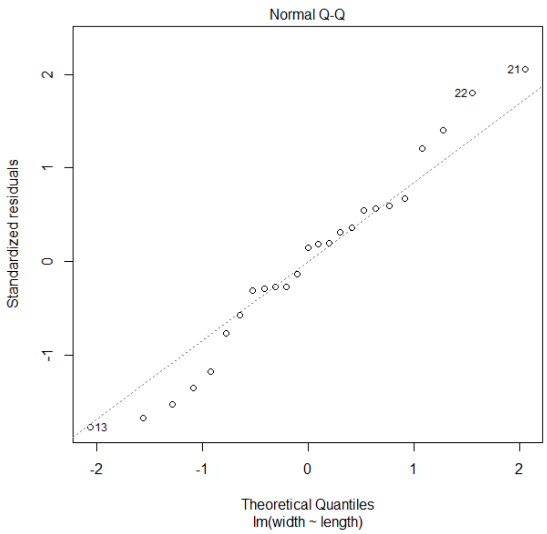
```
> crocs_plot <- ggplot(crocs, aes(x = length, y = width)) +
geom_jitter() + labs(x = "Mouth Length", y = "Mouth Width", title =
"Dimensions for 25 American Crocodiles") + geom_smooth(method = "lm",
se = FALSE)

> print(crocs_plot)
```



```
> crocs_lm <- lm(width ~ length, data = crocs)
> plot(crocs_lm, which = 1)
> plot(crocs_lm, which = 2)
```



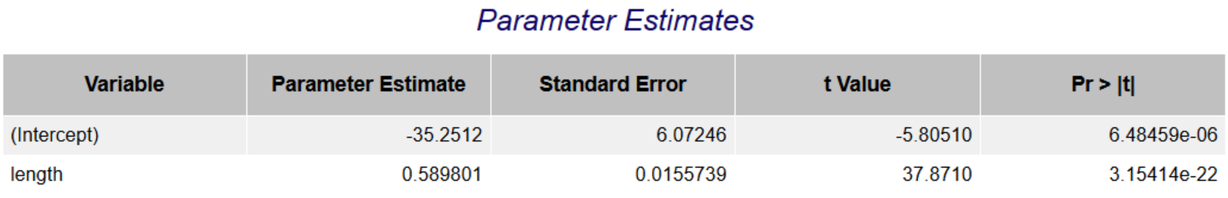


C) Are all assumptions met (or at least not terribly violated)? If not, explain the violation(s).

Assumptions might not be met, but they are not terribly violated. Since we are being asked to construct a confidence interval given an x-value in the dataset, we are probably okay to do this inference.

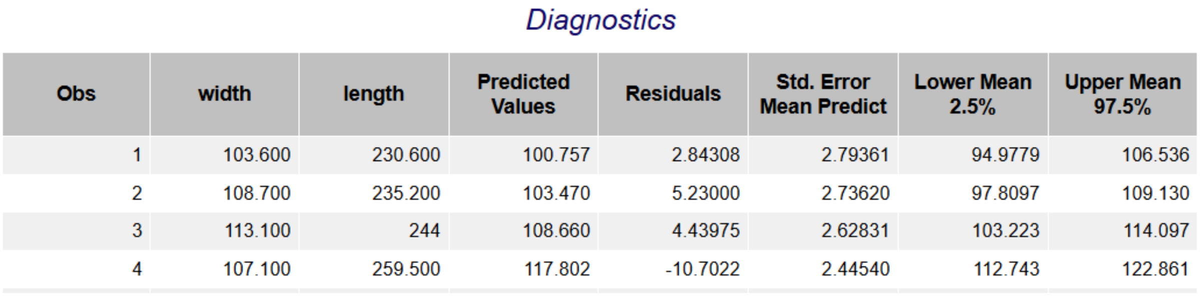
D) If assumptions are violated, stop here. If assumptions are okay, perform the inference:

Assumptions are met. We should perform the inference.



*Parameter Estimates*

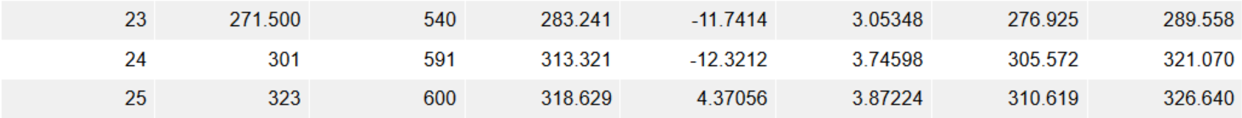
| Variable | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| (Intercept) | -35.2512 | 6.07246 | -5.80510 | 6.48459e-06 |
| length | 0.589801 | 0.0155739 | 37.8710 | 3.15414e-22 |



*Diagnostics*

| Obs | width | length | Predicted Values | Residuals | Std. Error Mean Predict | Lower Mean 2.5% | Upper Mean 97.5% |
|---|---|---|---|---|---|---|---|
| 1 | 103.600 | 230.600 | 100.757 | 2.84308 | 2.79361 | 94.9779 | 106.536 |
| 2 | 108.700 | 235.200 | 103.470 | 5.23000 | 2.73620 | 97.8097 | 109.130 |
| 3 | 113.100 | 244 | 108.660 | 4.43975 | 2.62831 | 103.223 | 114.097 |
| 4 | 107.100 | 259.500 | 117.802 | -10.7022 | 2.44540 | 112.743 | 122.861 |

…



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 23 | 271.500 | 540 | 283.241 | -11.7414 | 3.05348 | 276.925 | 289.558 |
| 24 | 301 | 591 | 313.321 | -12.3212 | 3.74598 | 305.572 | 321.070 |
| 25 | 323 | 600 | 318.629 | 4.37056 | 3.87224 | 310.619 | 326.640 |

The confidence interval corresponding to length = 600 is (310.62, 326.64). We are 95% confident that the population mean mouth width for crocodiles with mouth length 600 mm is between 310.62 and 326.64 mm.

Using R to perform the inference (I am not expecting you to do this with one line of code, but it helps to know how to do it!):

```
> predict(crocs_lm, newdata = data.frame(length = 600), interval =
"confidence", level = 0.95)

       fit      lwr      upr

1 318.6294 310.6191 326.6398
```

# Exam Problem 3

Maximum oxygen uptake volume ($VO_{2max}$, measured in mL per kg per minute) is a key indicator of aerobic fitness. A 2013 study suggested that a $VO_{2max}$ of 60 mL/kg/min is the minimum threshold necessary to play men's professional soccer at an elite level. The $VO_{2max}$ of severely injured athletes often decreases following injury.

The file ACL.csv contains the following variables, for 20 professional soccer players who underwent anterior cruciate ligament (ACL) reconstruction surgery:

- ID: the study ID number for the player
- Pre: the player's $VO_{2max}$ measured after ACL injury but before the surgery
- Post: the player's $VO_{2max}$ measured six months after the surgery

Assuming that this sample is representative of all players with ACL injuries requiring reconstruction surgery, estimate the amount by which an injured player's $VO_{2max}$ increases due to the surgery and six-month rehabilitation period.

A) What inferential procedure would you use to answer this question?

Best answer: matched pairs t confidence interval

0.75 pts: matched pairs t-test ("estimate" means we can't test a claim)

0.5 pts: any other t confidence interval

0.5 pts: inference in simple linear regression framework

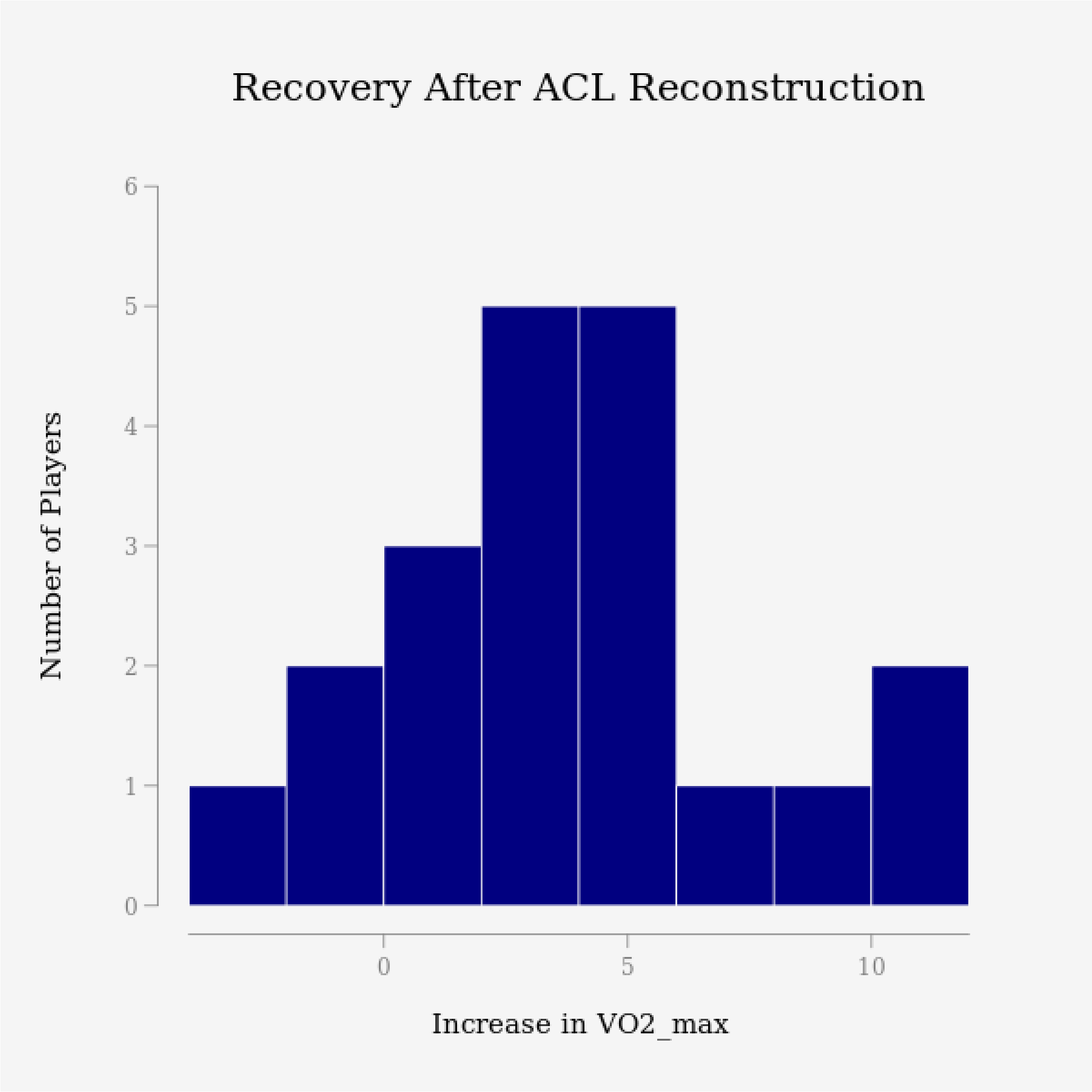
B) Perform background and exploratory data analysis to determine whether the assumptions of the procedure have been met. Write and paste your analysis below.

Check that a matched pairs t confidence interval is appropriate:

Do we have a simple random sample? Based on the information given, we probably don't, but the players are probably independent and assumed representative, so we're probably close enough for inference.

Do we know the population standard deviation of the difference? No, but we have a dataset that we can use to estimate it.

Do we meet the sample size rule of thumb?



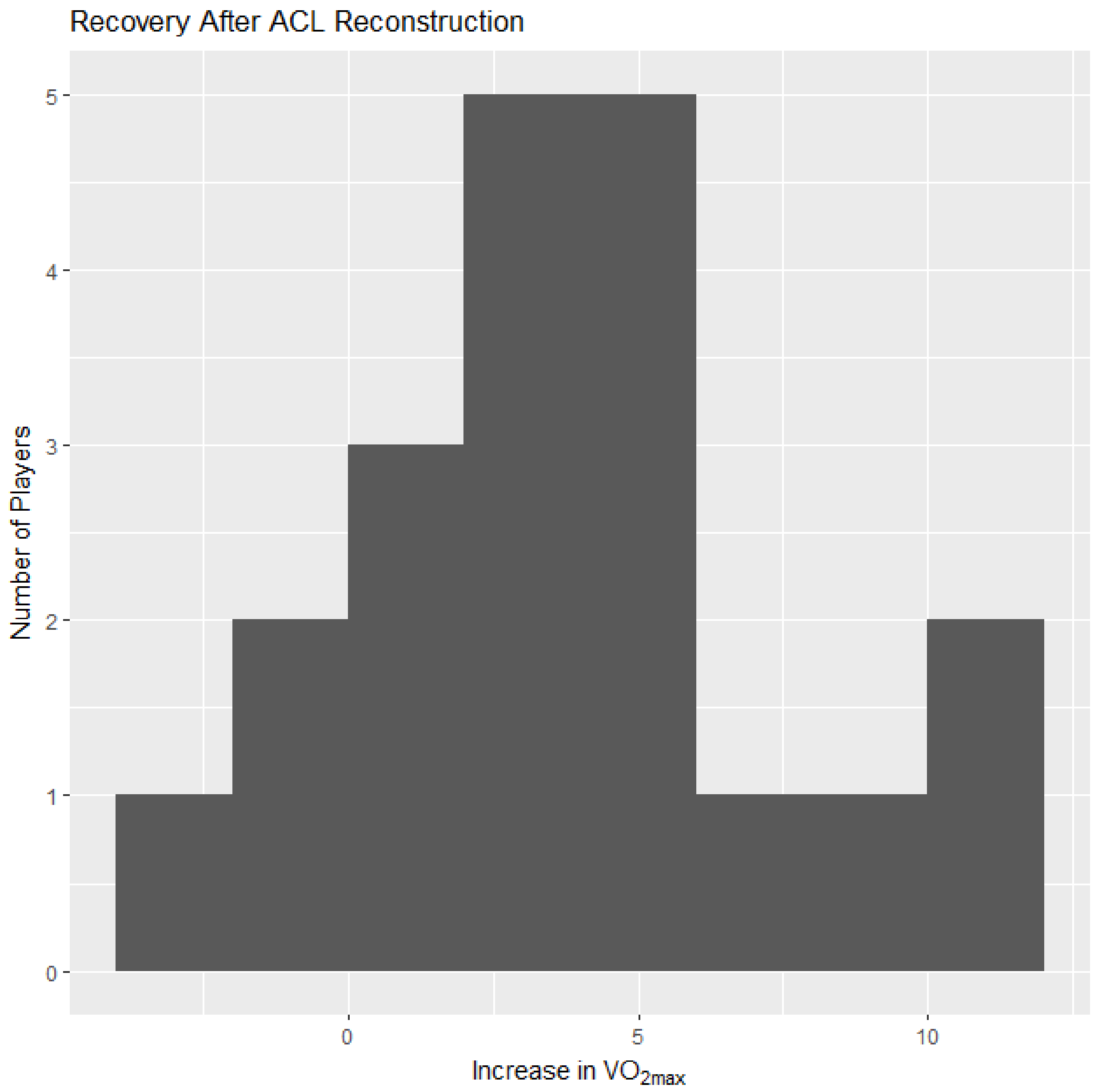
We have a sample size of 20 and the distribution of the difference is not terribly skewed. There do not appear to be outliers.

Using R to generate the histogram:

```
> ACL <- ACL %>% mutate(diff = Post - Pre)
```

```
> ACL_histogram <- ggplot(ACL, aes(x = diff)) + geom_histogram(center = 1, binwidth = 2) + labs(x = expression(paste("Increase in ", VO["2max"], sep = "")), y = "Number of Players", title = "Recovery After ACL Reconstruction")
```

```
> print(ACL_histogram)
```

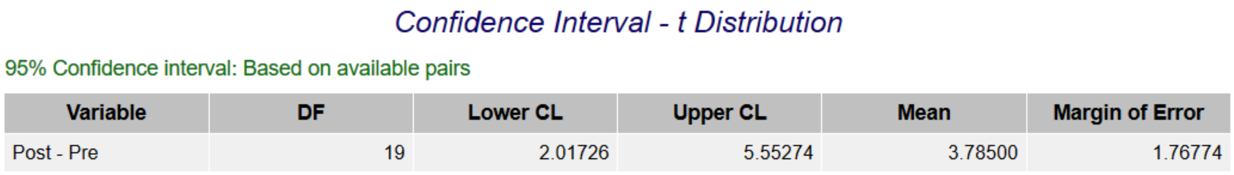


C) Are all assumptions met (or at least not terribly violated)? If not, explain the violation(s).

Assumptions are not terribly violated. We should proceed with inference.

D) If assumptions are violated, stop here. If assumptions are okay, perform the inference:

Assumptions are met. We should perform the inference.



*Confidence Interval - t Distribution*

95% Confidence interval: Based on available pairs

| Variable | DF | Lower CL | Upper CL | Mean | Margin of Error |
|---|---|---|---|---|---|
| Post - Pre | 19 | 2.01726 | 5.55274 | 3.78500 | 1.76774 |

The confidence interval is (2.017, 5.553). We are 95% confident that, on average, soccer players' VO$_{2max}$ increases by between 2.017 and 5.553 mL/kg/min following ACL reconstruction surgery and six months of rehabilitation.

Using R:

```
> t.test(ACL$diff, conf.level = 0.95)


        One Sample t-test


data:  ACL$diff
t = 4.4815, df = 19, p-value = 0.0002557
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.017255 5.552745
sample estimates:
mean of x
    3.785
```

A couple of alternative ways to do this:

```
> t.test(ACL$Post, ACL$Pre, paired = TRUE)


        Paired t-test


data:  ACL$Post and ACL$Pre
t = 4.4815, df = 19, p-value = 0.0002557
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.017255 5.552745
sample estimates:
mean of the differences
                  3.785


> with(ACL, t.test(Post, Pre, paired = TRUE))  # the super-fancy
version that doesn't require the $ sign


        Paired t-test


data:  Post and Pre
t = 4.4815, df = 19, p-value = 0.0002557
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.017255 5.552745
sample estimates:
mean of the differences
                  3.785
```