Section 1

Announcements & Recap

Section 2

Learning Objectives

R Learning Objectives

1. Create a variable consisting of the differences in corresponding observations.
2. Interpret R output providing confidence intervals and hypothesis tests for inference on the mean of a population of differences.
3. Interpret R output providing confidence intervals and hypothesis tests for inference on the difference in two population means.
4. Create histograms of just one group in order to check the normality conditions.

Statistical Learning Objectives

1. Continue discussing quantitative data, this week in regards to a mean of differences (paired data) and a difference in two means scenario.
2. Understand whether data is considered paired or from two independent samples.
3. Understand confidence intervals and hypothesis test for a paired mean of differences.
4. Understand confidence intervals and hypothesis test for a difference in two means.

Functions Covered in this Lab

1. `pt()`
2. `qt()`
3. `plot_t()` (custom)
4. `t.test()`

Section 3

Lab Tutorial

Collecting Two Sets of Numeric Data

This week, we're talking about inference when data from 2 **sets** of a *numeric* variable is collected. Let's first revisit the conditions.

Condition 1: Independence Within the Sample(s)

We are **always** hoping that the observations within the sample (or within the sample**s**) are independent of one another. When we have taken a random sample (or sample**s**), we can generalize to the appropriate population (or population**s**).

Condition 2: Independence Between the Samples

The question we have to ask ourselves, when dealing with 2 sets of observations of a numeric variable:

Are the two sets of observations independent of **one another**, such that observations in one sample tell us nothing about the observations in the other sample (and vice versa)?

If the answer is **yes**, then the appropriate method of inference is to keep the data separate and discuss **a difference in two population means**, with parameter $\mu_1 - \mu_2$.

If the answer is **no**, then the appropriate method of inference is to pair the data and discuss **the mean of the population of differences**, with parameter μ_{d}.

Paired Data, Mean of the Differences

For the *paired mean of the differences* scenario, we have

parameter μ_d

statistic $\bar{x}_d$

where the d's represent the difference in observations.

The conditions are

1. **Independence** (within the sample): The observed differences should be independent of one another. This condition is typically satisfied with a random sample of the differences.
2. **Normality**: The observed differences in observations should come from a (nearly-)normal population of differences. We will check by making a histogram of the sampled differences. We can relax the condition more and more for larger and larger sample sizes.

Births from North Carolina in 2004

The `penguins` data set doesn't contain anything that is paired data. So we turn our attention to a data set called `births.csv`.

This data set was collected by taking a random sample of 800 entries in a database containing information about babies born to cisgender heterosexual couples in North Carolina in 2004. Variables include the father's age, the mother's age, the weight of the baby, etc.

Let's say we want to estimate the average difference in the age of the father and the mother of babies born in North Carolina in 2004. Do you think that the father's age and the mother's age in this data set are independent of each other, or, do you think that the father's age and the mother's age in this data set are dependent on each other?

Paired Data

We should **pair** the data for father's age and mother's age together, as it was collected from the same observation (here, same baby). So we have one sample from one population of babies born in North Carolina in 2004.

Thus our parameter of interest is μ_{d}, which represents the population mean of the difference in the age of the father and the mother of a baby born in North Carolina in 2004.

Normality Condition

Recall that our normality condition requires that the observations of the difference in age need to come from a (nearly-)normal population.

So we will need to create a histogram of the **differences** in the parents' ages from the sample, and see if the distribution is approximately bell-shaped. ## Creating a Variable of the Differences In order to make this histogram, we need to create a new variable that is defined as the difference in the father's age and mother's age for the babies born in North Carolina in 2004.

First, let's read in the data. Run the `tryit1` code chunk in your notes file.

```
babies <- read.csv("births.csv", stringsAsFactors = TRUE)
```

Next, let's add a variable called `ageDiff` to this `babies` data set. We will define the difference as father's age (`fage`) minus mother's age (`mage`). Try this out in the `tryit2` code chunk in your notes file.

```
babies$ageDiff <- babies$fage - babies$mage
```

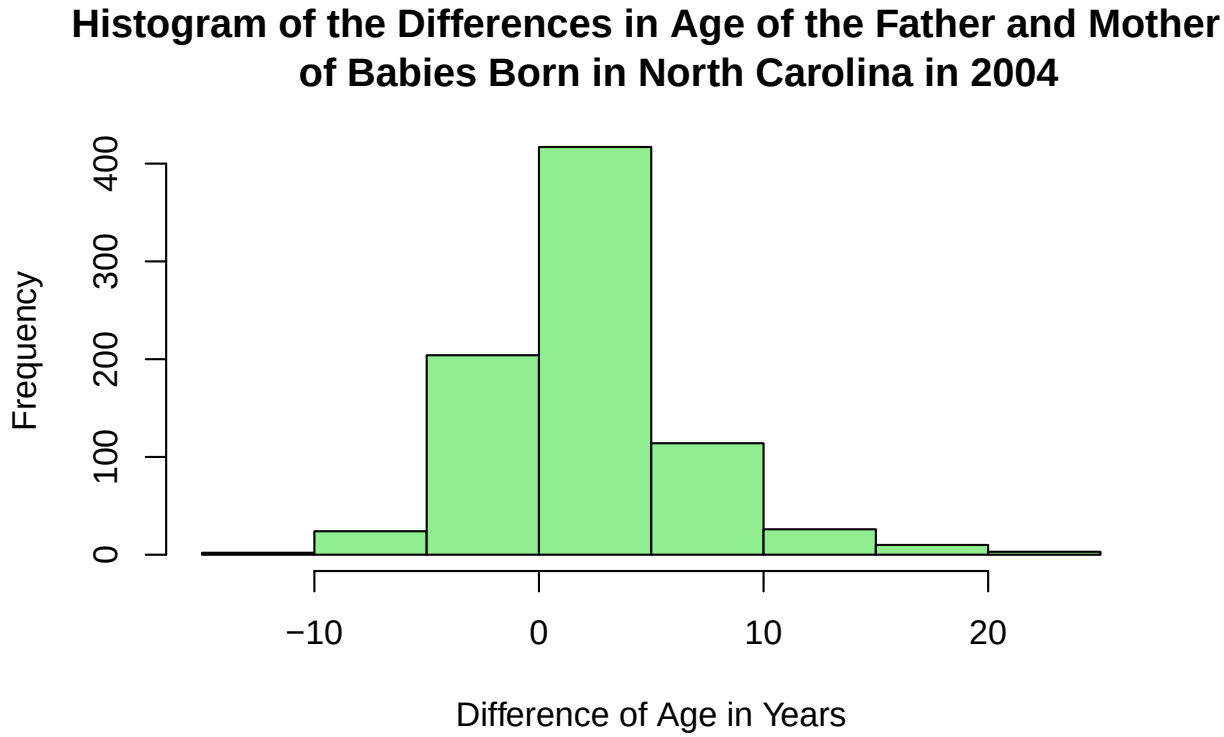
Creating a Histogram of the Differences

We can now create the histogram of the differences in age between the father and the mother for the sampled babies born in North Carolina in 2004.

The title is really long, so we will go to the next line where we want a line break. Try this in `tryit3`.

```
hist(babies$ageDiff,
     main = "Histogram of the Differences in Age of the Father and Mother
     of Babies Born in North Carolina in 2004",
     xlab = "Difference of Age in Years",
     col = "lightgreen")
```

Histogram of the Differences



This histogram is unimodal and slightly right skewed. However, the sample size (800) is definitely large enough to relax the normality condition.

(Remember that the reason we can relax the normality condition is that the sample mean of the differences $\bar{x}_d$ will be approximately normal.)

Estimating the Mean of the Differences

We want to estimate the mean of the differences in age between fathers and mothers for all babies born in North Carolina in 2004. This requires creating a confidence interval. Let's create a 90% confidence interval. Try this out in the `tryit4` code chunk in your notes file.

```
t.test(babies$ageDiff,
       conf.level = 0.90)
```

```
##
##  One Sample t-test
##
## data:  babies$ageDiff
## t = 17.265, df = 799, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  2.380281 2.882219
## sample estimates:
## mean of x
##   2.63125
```

Hypothesis Test for a Population Mean of the Differences

Let's see if we have evidence for the claim that, on average, fathers are older than mothers, for babies born in North Carolina in 2004.

Thus we are testing the hypotheses:

$H_0 : \mu_{\text{d}} = 0$

$H_A : \mu_{\text{d}} > 0$

where μ_{d} is the mean of the differences in the father's age and mother's age (father minus mother) of babies born in North Carolina in 2004.

Code for the Hypothesis Test

Try this code out in the `tryit5` code chunk in your notes file.

```
t.test(babies$ageDiff,
       mu = 0,
       alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  babies$ageDiff
## t = 17.265, df = 799, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  2.380281      Inf
## sample estimates:
## mean of x
##   2.63125
```

Conclusion

From the R output, we see that the test statistic is $t = 17.265$ and the p-value is tiny (much less than any of the standard significant levels.) We reject the null hypothesis.

Our analysis suggests that the mean of the differences in the father's age and mother's age (father minus mother) of babies born in North Carolina in 2004 is greater than zero.

Difference in Two Means

When we have two independent samples, we need to work with the *difference in two means*. In this scenario, we have

parameter $\mu_1 - \mu_2$

statistic $\bar{x}_1 - \bar{x}_2$

for group 1 and group 2

Difference in Two Means

The conditions are

1. **Independence within each sample:** The observations within each sample are independent. Typically satisfied by taking two random samples (one from each population) or by taking one random sample and splitting it into two independent groups (e.g., in-state students and out-of-state students).
2. **Independence between the samples:** The two samples are independent of one another such that observations in one sample tell us nothing about the observations in the other sample (and vice versa).
3. **Normality** for group 1: The sample observations from group 1 should come from a nearly-normal population. We will check by making a histogram of the sample observations from group 1. We can relax the condition more and more for larger and larger sample sizes.
4. **Normality** for group 2: The sample observations from group 2 should come from a nearly-normal population. We will check by making a histogram of the sample observations from group 2. We can relax the

Back to the Penguins!

Run the `tryIt6` code chunk, to read in the `penguins` data.

```
penguins <- read.csv("penguins.csv", stringsAsFactors = TRUE)
```

Let's compare the mean flipper lengths of Adelie and Chinstrap penguins.

Subsetting the Data to Only Include Two Groups

First, we'll subset the data to just involve Adelie and Chinstrap penguins. This is only because we're not interested in Gentoo penguins for this question, so we'll take them out.

Since we are interested in penguins that are *either* Adelie *or* Chinstrap, R has an easy way to achieve this by using the `%in%` operator. Don't forget the double quotes around Adelie and Chinstrap!

Run the code in the `tryit7` code chunk in your notes document.

```
penguinsSubset <- subset(penguins,
                         species %in% c("Adelie", "Chinstrap"))
```

Normality Condition x 2

To check the normality conditions, we will need to make **two** histograms: one for the Adelie penguins, and another for the Chinstrap penguins.

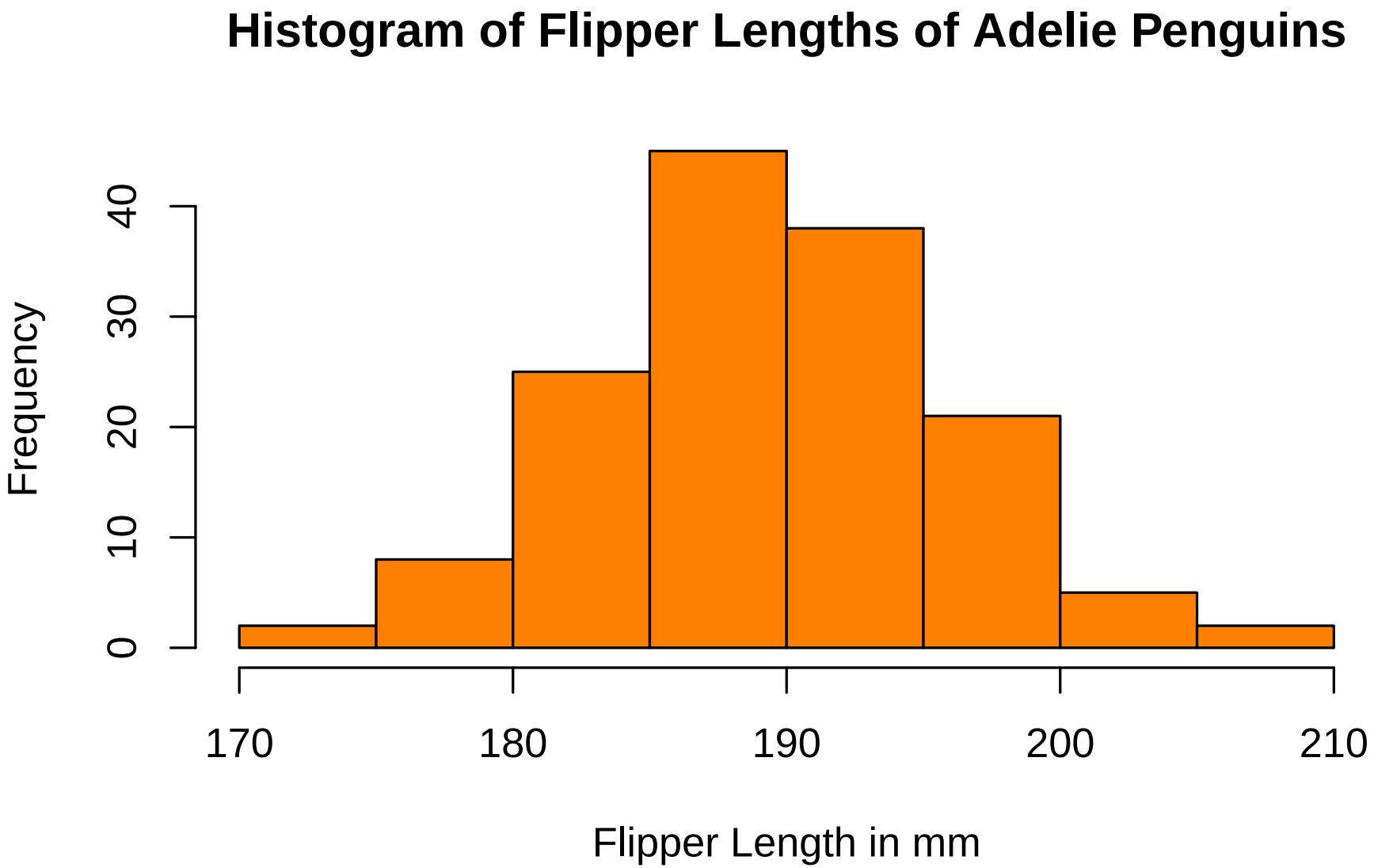
Since we only need these one group subsets for the histograms, we will not bother to give it a name. Instead, we will embed the subset code inside the histogram code.

Histogram of Just Adelie Penguins Flipper Length

Run this code in the `tryit8` code chunk in your notes document.

```
hist(penguinsSubset$flipper_length_mm[penguinsSubset$species == "Adelie"],
     main = "Histogram of Flipper Lengths of Adelie Penguins",
     xlab = "Flipper Length in mm",
     col = "darkorange1")
```

Histogram of Just Adelie Penguins Flipper Length

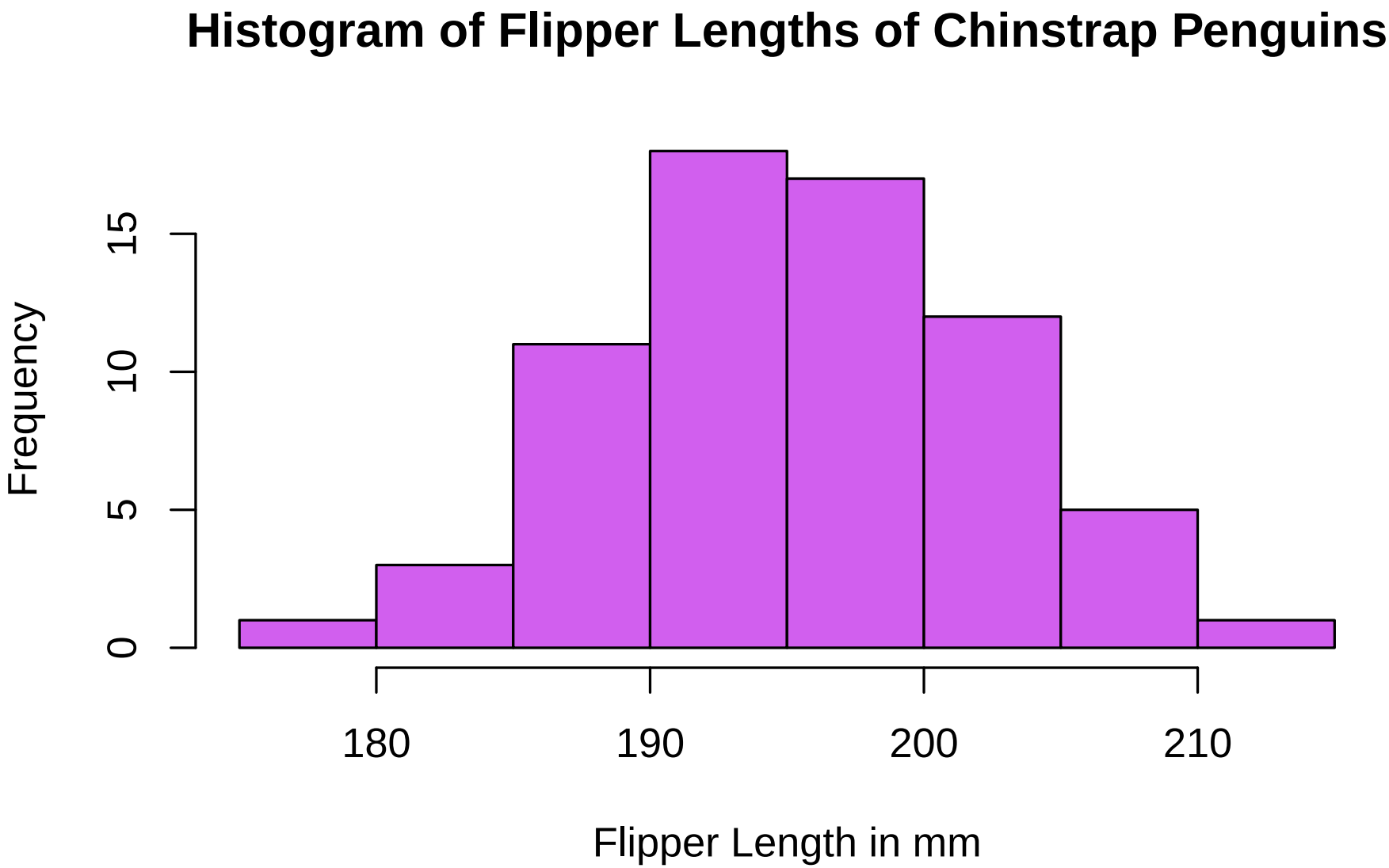


Histogram of Just Chinstrap Penguins Flipper Length

Then run this code in the `tryit9` code chunk in your notes document.

```
hist(penguinsSubset$flipper_length_mm[penguinsSubset$species == "Chinstrap"],
     main = "Histogram of Flipper Lengths of Chinstrap Penguins",
     xlab = "Flipper Length in mm",
     col = "mediumorchid2")
```

Histogram of Just Chinstrap Penguins Flipper Length



Normality Condition

Each of the two histograms appear to be unimodal and approximately bell-shaped, so it is reasonable that the flipper lengths come from populations with normal distributions, and the normality condition is reasonably met for both groups.

(Note that we don't need to check the sample sizes to see if we can relax the normality condition because we think the normality condition is met.)

Hypothesis Test for the Difference in Two Means

We would like to test the claim that there **is** a difference in the mean flipper length for the two groups, Adelie penguins and Chinstrap penguins.

$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0$

The parameter here is $\mu_1 - \mu_2$, where μ_1 is the mean flipper length in mm for Adelie penguins on Palmer Archipelago, and μ_2 is the mean flipper length in mm for Chinstrap penguins on Palmer Archipelago.

Performing the *t*-test for the Difference in Two Population Means

Try this code out in the `tryit10` code chunk in your notes document.

```
t.test(flipper_length_mm ~ species,
       data = penguinsSubset,
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  flipper_length_mm by species
## t = -5.6115, df = 120.88, p-value = 1.297e-07
## alternative hypothesis: true difference in means between group Adelie and group C
## 95 percent confidence interval:
##  -7.739129 -3.702450
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                190.1027                 195.8235
```

Conclusion

From the R output, we see that the test statistic is $t = -5.6115$ and the p-value is tiny (much less than any of the standard significance levels.) We reject the null hypothesis.

Our analysis suggests that there is a difference between the mean flipper length in mm for Adelie penguins on Palmer Archipelago and the mean flipper length in mm for Chinstrap penguins on Palmer Archipelago.

Difference in Two Means Confidence Interval

When we have statistically significant results that suggest there is a difference between the two populations means, it is helpful to construct a confidence interval to find a range of reasonable values for that difference. (Confidence intervals are also helpful when we are simply interested in estimating the parameter.)

As a reminder, the parameter here is $\mu_1 - \mu_2$, where μ_1 is the mean flipper length in mm for Adelie penguins on Palmer Archipelago, and μ_2 is the mean flipper length in mm for Chinstrap penguins on Palmer Archipelago.

Computing the 95% Confidence Interval for the Difference in Two Population Means

Try this code in the `tryit11` code chunk in your notes document.

```
t.test(flipper_length_mm ~ species,
       data = penguinsSubset,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  flipper_length_mm by species
## t = -5.6115, df = 120.88, p-value = 1.297e-07
## alternative hypothesis: true difference in means between group Adelie and group C
## 95 percent confidence interval:
##  -7.739129 -3.702450
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                190.1027                 195.8235
```

Interpreting the Confidence Interval

The 95% confidence interval is (-7.739, -3.702). A fairly standard way to interpret this confidence interval is to say

"We estimate, with 95% confidence, that the difference between the mean flipper length in mm for Adelie penguins on Palmer Archipelago and the mean flipper length in mm for Chinstrap penguins on Palmer Archipelago is between -7.739 mm and -3.702."

This confidence interval is awkward. A better way to understand what the confidence interval tells us is to say

"We estimate, with 95% confidence, that the mean flipper length in mm for Adelie penguins on Palmer Archipelago is between 3.702 mm and 7.739 mm *less* than the mean flipper length in mm for Chinstrap penguins on Palmer Archipelago."