# PSTAT 5LS Lab 5

TA NAME HERE

Spring 2025

Section 1

## Announcements & Recap

Section 2

# Learning Objectives

# R Learning Objectives

1. Create histograms for just one group in order to check the normality conditions.
2. Generate and interpret R output providing confidence intervals and hypothesis tests for inference for two independent means $\mu_1 - \mu_2$.
3. Create a variable consisting of the differences in corresponding observations.
4. Generate and interpret R output providing confidence intervals and hypothesis tests for inference for and the mean of the differences $\mu_d$.

# Statistical Learning Objectives

1. Differentiate between scenarios for one mean, two independent means, and the mean of the differences.
2. Understand confidence intervals and hypothesis tests for a difference in two independent means.
3. Understand confidence intervals and hypothesis tests for the mean of the differences.

Section 3

# Lab Tutorial Part 1

# Collecting Two Sets of Numeric Data

Today, 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 $\mu_{\text{d}}$.

# Paired or Independent?

Consider the following scenarios and determine if the data are paired or if the data come from independent samples.

a. A company wants to compare job satisfaction levels between their full-time and part-time employees. They survey 100 full-time and 100 part-time employees.

b. A psychologist is studying the effects of caffeine on reaction times. They test each participant's reaction time once after drinking regular coffee and once after drinking decaf coffee, with tests done on separate days.

# Paired or Independent?

c. An educational researcher wants to investigate if there's a difference in math performance between students who attend public schools and those who attend private schools. They randomly select 200 8th grade students from public schools and 200 from private schools, then administer the same standardized math test to both groups.

d. A language learning app developer wants to assess the improvement in users' vocabulary after using the app for two months. They test the vocabulary of 100 users when they first download the app and then test the users again after two months of app usage.

# Difference in Two Means

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

$$\mu_1 - \mu_2$$

where $\mu_1$ represents the population mean for "group 1" and $\mu_2$ represents the population mean for "group 2".

The point estimate is

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

# Difference in Two Means

The conditions are

1. **Independence within each sample:** The observations within each sample are independent.

This condition is 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** (must check for both groups): Each sample of observations should come from a nearly-normal population.

We will check by making a histogram of the sample observations from each of the groups. We can relax the condition more and more for larger and larger sample sizes.

# Back to the Penguins!

Run the `loadPenguins` code chunk in your notes so that we can use the penguins data set.

# Subsetting the Data to Only Include Two Groups

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

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 `tryit1` 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.

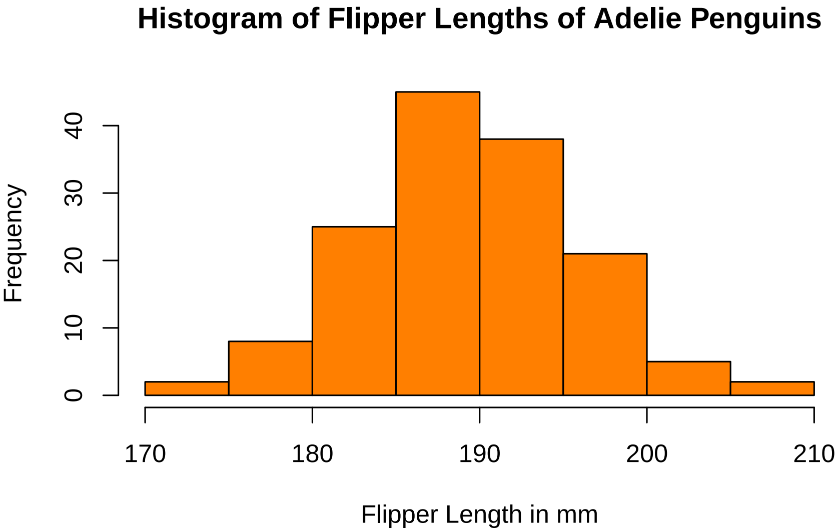
Since we only need these one group subsets for the histograms, we will not bother to give it a name. Instead, we will embed the subset code inside the histogram code.

# Histogram of Just Adelie Penguins Flipper Length

Let's first create a histogram of the flipper lengths for the Adelie penguins. Run the `tryit2` code chunk in your notes document to generate this histogram.

```r
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

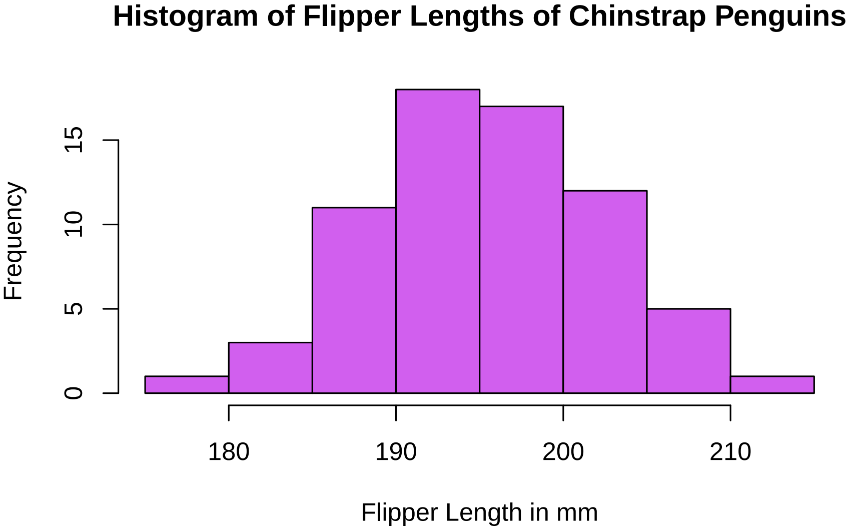


# Histogram of Just Chinstrap Penguins Flipper Length

Next up, let's create a histogram of the flipper lengths for the Chinstrap penguins. Change the code in the `tryit3` code chunk in your notes document so that you get the histogram we need.

```r
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

Does the normality condition appear to be satisfied?

# 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 $\mu_1$ is the mean flipper length in mm for Adelie penguins on Palmer Archipelago, and $\mu_2$ is the mean flipper length in mm for Chinstrap penguins on Palmer Archipelago.

# Performing the $t$-test for the Difference in Two Population Means

The code to conduct the two-sample test is in the `tryit4` code chunk in your notes document. Make sure to specify the alternative in the `alternative` argument ("two.sided", "less", "greater") before running the chunk.

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

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

```
##
##  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 A
## 95 percent confidence interval:
##  -7.739129 -3.702450
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                190.1027                195.8235
```

# Decision and 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 $\mu_1$ is the mean flipper length in mm for Adelie penguins on Palmer Archipelago, and $\mu_2$ is the mean flipper length in mm for Chinstrap penguins on Palmer Archipelago.

# Computing the 98% Confidence Interval for the Difference in Two Population Means

Try this code in the tryit5 code chunk in your notes document. Be sure to specify the confidence level before running the chunk.

```
t.test(flipper_length_mm ~ species,
       data = penguinsSubset,
       conf.level = 0.98)
```

# Computing the 98% Confidence Interval for the Difference in Two Population Means

```
##
##  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 gr
## 98 percent confidence interval:
##  -8.124296 -3.317284
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                190.1027                195.8235
```

# Interpreting the Confidence Interval

The 98% confidence interval is (-8.124, -3.317).

A fairly standard way to interpret this confidence interval is to talk about reasonable values of the difference between the two means:

"We estimate, with 98% 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 -8.124 mm and -3.317."

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

"We estimate, with 98% confidence, that the mean flipper length in mm for Adelie penguins on Palmer Archipelago is between 3.317 mm and 8.124 mm *less* than the mean flipper length in mm for Chinstrap penguins on Palmer Archipelago."

# Tie between Hypothesis Tests and Confidence Intervals

Remember that confidence intervals use what we have from our sample(s) to give us a range of values we think are reasonable for the parameter.

If we had not already conducted a hypothesis test for the difference in the two means, the 98% confidence interval we created could be used to test the hypotheses

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

at the $\alpha = 0.02$ significance level.

# Tie between Hypothesis Tests and Confidence Intervals

The 98% confidence interval for the difference between the mean flipper length in mm for Adelie penguins and the mean flipper length in mm for Chinstrap penguins is -8.124 mm to -3.317. Do we reject or fail to reject $H_0 : \mu_1 - \mu_2 = 0$ at the $\alpha = 0.02$ significance level?

Section 4

# Lab Tutorial Part 2

# Mean of the Differences for Paired Data

When we are working with paired data, the parameter is the *mean of the differences*

$$\mu_{\mathrm{d}}$$

The point estimate (sample statistic) is

$$\bar{x}_{\mathrm{d}}$$

where the d's represent the difference in observations.

# Mean of the Differences for Paired Data

The conditions are

1. **Independence**: The observed differences should be independent of one another. This condition is typically satisfied with a random sample of the differences, but you might need to think about whether the differences are independent from observation to observation.
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 differences in the sample. If the sample looks like it came from a population that is normally distributed, then this condition is met. Note that we can relax the condition more and more for larger and larger sample sizes.

# Births from North Carolina in 2004

The penguins data set we have used 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 both ages were 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 $\mu_{\mathsf{d}}$, 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 `tryit6` code chunk in your notes file.

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

# Creating a Variable of the Differences

Let's look at structure of the babies data set so we can figure out which variables indicate the father's age and the mother's age.

Write the code you need to peek at the structure of the data set in the tryit7 code chunk and then run the code chunk.

```
str(babies)
```

What are the variables that indicate the father's age and the mother's age?

# Creating a Variable of the Differences

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`).

Create this new variable in the `tryit8` code chunk in your notes file.

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

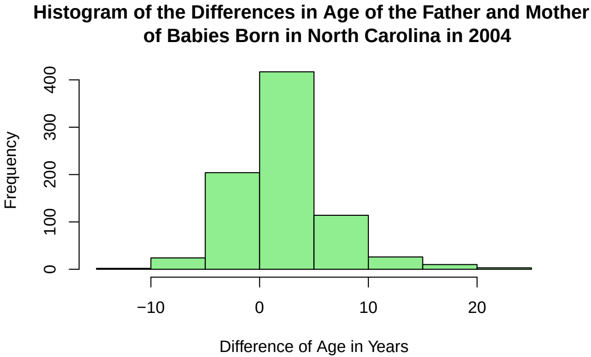
# Creating a Histogram of the Differences

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. Specify the variable name in the `tryit9` code chunk in your notes and then run it to create a histogram.

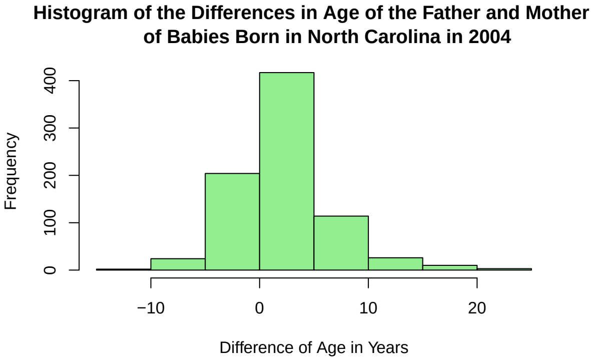
```
hist(babies$ageDiff,
     main = "Histogram of the Differences in Age of the Father and Mother
     of Babies Born in North Carolina in 2004",
     xlab = "Difference of Age in Years",
     col = "lightgreen")
```

# Histogram of the Differences



Does the normality condition appear to be satisfied?

# 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 `tryit10` code chunk in your notes file.

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

# Confidence Interval R Output

```r
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
```

# Interpreting the Confidence Interval

Interpret the confidence interval.

# Interpreting the Confidence Interval

The 90% confidence interval suggests that, on average, fathers are between 2.38 and 2.88 years older than mothers for all babies born in North Carolina in 2004.