STATS 250 Lab 12

Paired Data and Difference of Two Means

Nick Seewald
nseewald@umich.edu

Week of 11/16/2020

Reminders

Your tasks for the week running Friday 11/13 - Friday 11/20

Task	Due Date	Submission
M-Write 2 Revision	Thursday 11/19 4:59PM ET	Canvas
Lab 12	Friday 11/20 8:00AM ET	Canvas
Homework 9	Friday 11/20 8:00AM ET	course.work

M-Write Office Hours on Canvas!

Homework 8 Comments

Question 3b:

The EPA claims that a 2012 Prius gets 50 MPG (city and highway mileage combined). Do these data provide strong evidence against this estimate for drivers who participate on fueleconomy.gov?

Make sure to state your hypotheses, check the conditions, calculate the test statistic, determine the p-value, evaluate the p-value and the compatibility of the null model, and make a conclusion in the context of the problem (and, if necessary, make a recommendation).

Homework 8 Comments

Question 3b:

> The EPA claims that a 2012 Prius gets 50 MPG (city and highway mileage combined). Do these data provide strong evidence against this estimate for drivers who participate on fueleconomy.gov?

- This is a question about a **mean**, not a proportion: inference is on μ, not p.
- $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$; order matters here.
- Two-sided p-value: double `pt()` output!
- SHOW WORK; conclusion IN CONTEXT; check ALL conditions

Homework 8 Comments

Question 3c:

Calculate a 95% confidence interval for the average gas mileage of a 2012 Prius by drivers who participate on fueleconomy.gov.

- Make sure to use the correct t^* value:
  - $n = 14$, so $\text{df} = 14 - 1 = 13$
  - $t^* =$ `qt(p = 0.975, df = 13)` $= 2.16$

Homework 8 Comments

Question 6d:

Drive-thru window. Calculate the effect size for this hypothesis test.

$$d = \frac{\mu - \mu_0}{\sigma}$$

We don't know μ or σ! So we **estimate** d using $\hat{d}$:

$$\hat{d} = \frac{\bar{x} - \mu_0}{s}$$

Again, **order matters**.

Homework 8 Comments

Question 8: Type 1 and Type 2 errors

	Decide in favor of H_0	Decide in favor of H_A
H_0 true	✓	✗ Type 1 error
H_A true	✗ Type 2 Error	✓

H_0: The RC airplane's landing gear is down; the plane is cleared to land

H_A: The RC airplane's landing gear is not down; the plane is not cleared to land and will require troubleshooting

Name That Scenario 1

Do you get cookies with your dinner? A UM student decided to stop by MoJo one day at dinner, and asks each student to report the number of cookies they ate while at MoJo. From this data, an appropriate random sample will be selected. The UM student wishes to see if on average, the UM students ate more than one cookie at MoJo during their dinner.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Name That Scenario 1

Do you get cookies with your dinner? A UM student decided to stop by MoJo one day at dinner, and asks each student to report the number of cookies they ate while at MoJo. From this data, an appropriate random sample will be selected. The UM student wishes to see if on average, the UM students ate more than one cookie at MoJo during their dinner.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Name That Scenario 2

Does MoJo cure homesickness? A UM student is interested in seeing if perhaps cookies are just a fond memory for freshman students embarking on their first semester away from home. She gathers a large random sample of UM freshmen students, to ask whether or not they enjoy the cookies at MoJo. Next, she gathers a large random sample of UM students who are not freshmen, to ask whether or not they enjoy the cookies at MoJo. She hopes to compare the differences in these two rates.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Name That Scenario 2

Does MoJo cure homesickness? A UM student is interested in seeing if perhaps cookies are just a fond memory for freshman students embarking on their first semester away from home. She gathers a large random sample of UM freshmen students, to ask whether or not they enjoy the cookies at MoJo. Next, she gathers a large random sample of UM students who are not freshmen, to ask whether or not they enjoy the cookies at MoJo. She hopes to compare the differences in these two rates.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Name That Scenario 3

Are more cookies made at MoJo than at East Quad, on average? A UM student decides to select a random smaple of 30 days from the Winter 2019 semester. For each of those 30 days, they ask each dining hall to report the number of cookies baked. These results will be used to assess whether more cookies are made at MoJo than at East Quad, on average.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Name That Scenario 3

Are more cookies made at MoJo than at East Quad, on average? A UM student decides to select a random sample of 30 days from the Winter 2019 semester. For each of those 30 days, they ask each dining hall to report the number of cookies baked. These results will be used to assess whether more cookies are made at MoJo than at East Quad, on average.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Name That Scenario 4

Who has the tastiest cookies? A UM student decides to stop by MoJo to get 40 freshly baked cookies, then stops at East Quad for another 40 freshly baked cookies. She then gets a random sample of 40 UM freshmen, and has them each take a blind taste test. They will taste each cookie, one at a time, without knowing its origin, and select the cookie they like the most. The UM student would like to see if a majority pick MoJo cookies as the tastiest.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Name That Scenario 4

Who has the tastiest cookies? A UM student decides to stop by MoJo to get 40 freshly baked cookies, then stops at East Quad for another 40 freshly baked cookies. She then gets a random sample of 40 UM freshmen, and has them each take a blind taste test. They will taste each cookie, one at a time, without knowing its origin, and select the cookie they like the most. The UM student would like to see if a majority pick MoJo cookies as the tastiest.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Name That Scenario 5

Do athletes love cookies? As a follow up to a previous observation, a UM student decides that it might be best to gather data about whether the student is an athlete or not, as the number of cookies they eat in a week might differ, on average. The UM student gathers a large random sample of UM athletes to ask them to self-report the number of cookies they ate at MoJo the week before. Then the UM student gathers a large random sample of UM non-athletes to ask them to self-report the number of cookies they ate at MoJo the week before.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Name That Scenario 5

Do athletes love cookies? As a follow up to a previous observation, a UM student decides that it might be best to gather data about whether the student is an athlete or not, as the number of cookies they eat in a week might differ, on average. The UM student gathers a large random sample of UM athletes to ask them to self-report the number of cookies they ate at MoJo the week before. Then the UM student gathers a large random sample of UM non-athletes to ask them to self-report the number of cookies they ate at MoJo the week before.

a. One-sample Z-test for a population proportion

b. Two-sample Z-test for the comparison of two population proportions

c. One-sample t-test for a population mean

d. Paired t-test for a population mean of the difference

e. Two-sample t-test for the comparison of two population means

Paired Data (line ~115)

Are textbooks actually cheaper online? Let's compare prices of textbooks at the UCLA bookstore and Amazon for a random sample of 73 courses in the spring (winter) semester of 2010.

```
textbooks <- read.csv("textbooks.csv")
head(textbooks)
```

```
  dept_abbr course            isbn ucla_new amaz_new
1    Am Ind   C170 978-0803272620    27.67    27.95
2    Anthro      9 978-0030119194    40.59    31.14
3    Anthro   135T 978-0300080643    31.68    32.00
4    Anthro  191HB 978-0226206813    16.00    11.52
5   Art His  M102K 978-0892365999    18.95    14.21
6   Art His   118E 978-0394723693    14.95    10.17
```

Paired Data

- Natural correspondence between UCLA price and Amazon price: they're for the same book!
- Same "machinery" as a one-population mean t-test

Key Idea: When working with paired data, we'll work with *differences* between the paired observations. Our questions are about μ_{diff}, the average difference in the population.

$$t = \frac{\bar{x}_{\text{diff}} - \mu_0}{s_{\text{diff}}/\sqrt{n}}$$

Paired t-Test (line ~131)

- Same "machinery" as a one-population mean t-test, just using *differences*
- We need to make a variable that represents the differences!

```
names(textbooks)
```

```
[1] "dept_abbr" "course"    "isbn"      "ucla_new"  "amaz_new"
```

```
textbooks$diff <- ___________ - ___________
```

Paired *t*-Test (line ~131)

- Same "machinery" as a one-population mean *t*-test, just using *differences*
- We need to make a variable that represents the differences!

```
names(textbooks)
```

```
[1] "dept_abbr" "course"    "isbn"      "ucla_new"  "amaz_new"
```

```
textbooks$diff <- ___________ - ___________
```

```
textbooks$diff <- textbooks$ucla_new - textbooks$amaz_new
head(textbooks)
```

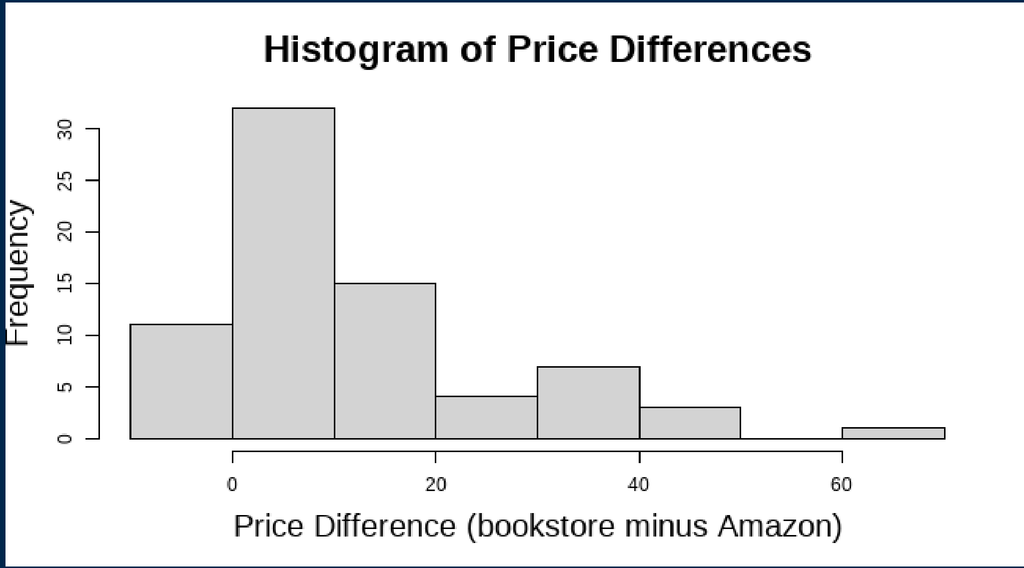
```
  dept_abbr course            isbn ucla_new amaz_new  diff
1    Am Ind   C170 978-0803272620    27.67    27.95 -0.28
2    Anthro      9 978-0030119194    40.59    31.14  9.45
3    Anthro   135T 978-0300080643    31.68    32.00 -0.32
4    Anthro  191HB 978-0226206813    16.00    11.52  4.48
5   Art His  M102K 978-0892365999    18.95    14.21  4.74
6   Art His   118E 978-0394723693    14.95    10.17  4.78
```

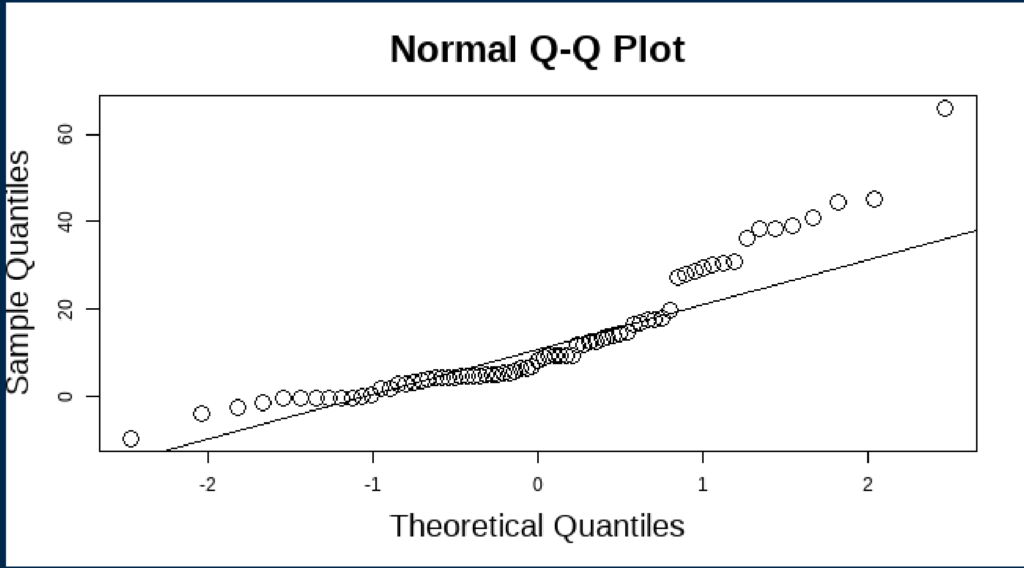
Paired t-Test: Check Conditions! (line ~145)

- Same "machinery" as a one-population mean t-test, just using *differences*

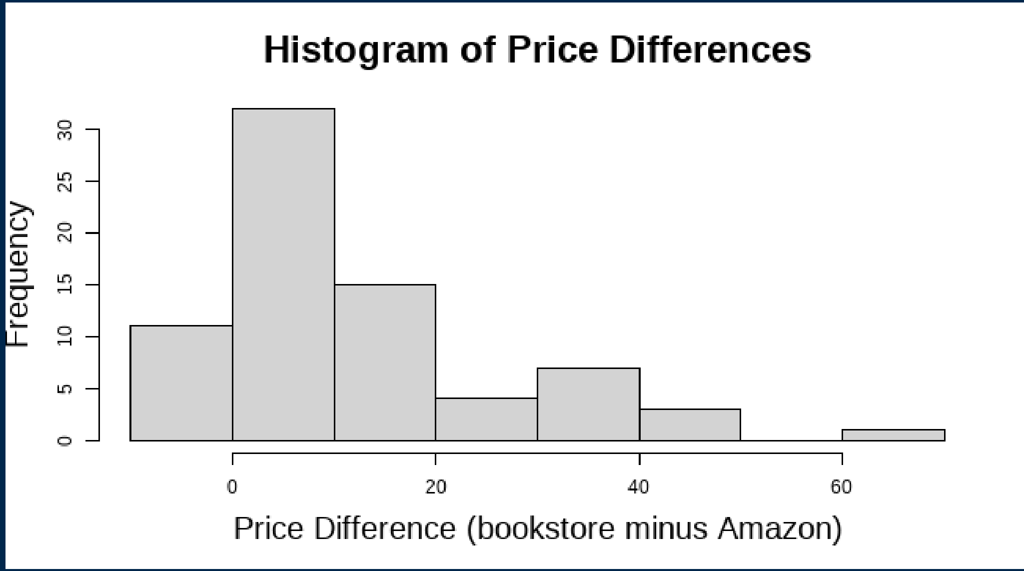
```
hist(textbooks$diff, main = "Histogram of Price
     xlab = "Price Difference (bookstore minus
```

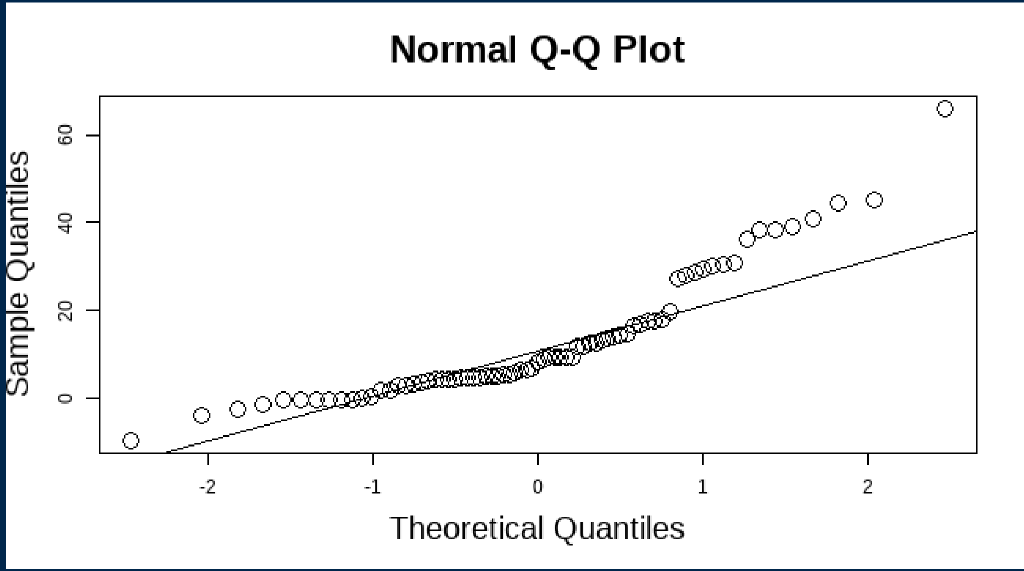
```
qqnorm(textbooks$diff)
qqline(textbooks$diff)
```





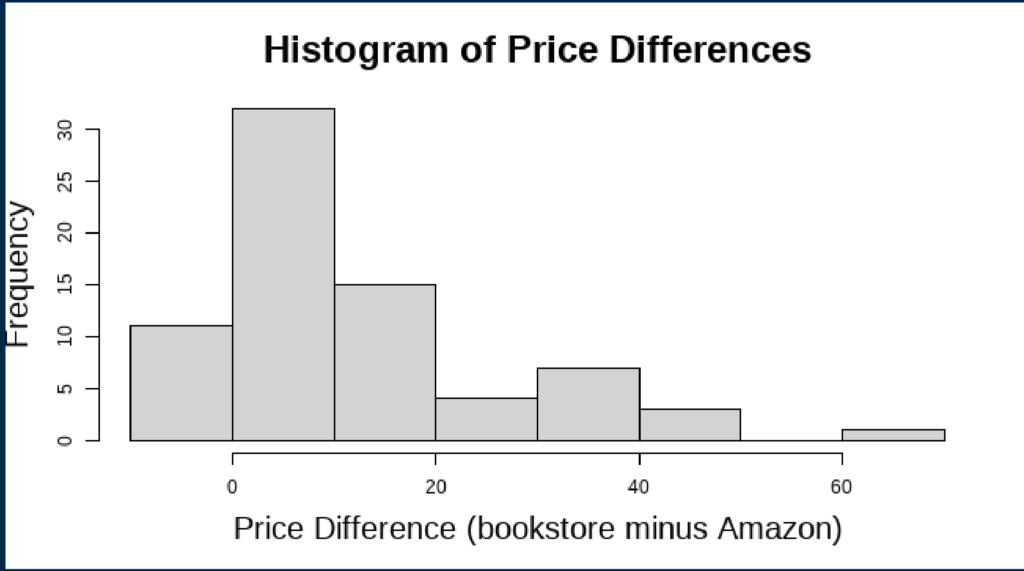
Paired t-Test: Check Conditions! (line ~145)

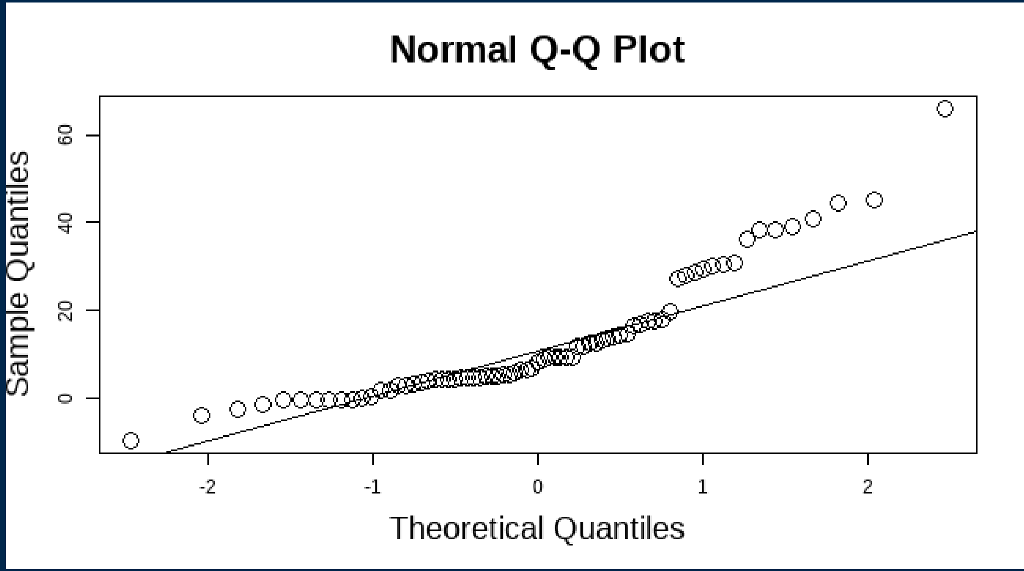




Do the differences seem to come from a normally-distributed population?

Paired t-Test: Check Conditions! (line ~145)





Do the differences seem to come from a normally-distributed population?

NOPE. But, there are 73 of them, so we can use the central limit theorem to say $\bar{x}_{\text{diff}}$ is nearly normal, which is good enough.

Paired t-Test (line ~157)

We want to know if there's a *difference* between the prices, on average.

$$H_0 : \mu_{\text{diff}} = 0 \quad \text{vs.} \quad H_a : \mu_{\text{diff}} \neq 0,$$

Same "machinery" as a one-population mean t-test, just using *differences*.

```
t.test(textbooks$diff, mu = 0, alternative = "two.sided")
```

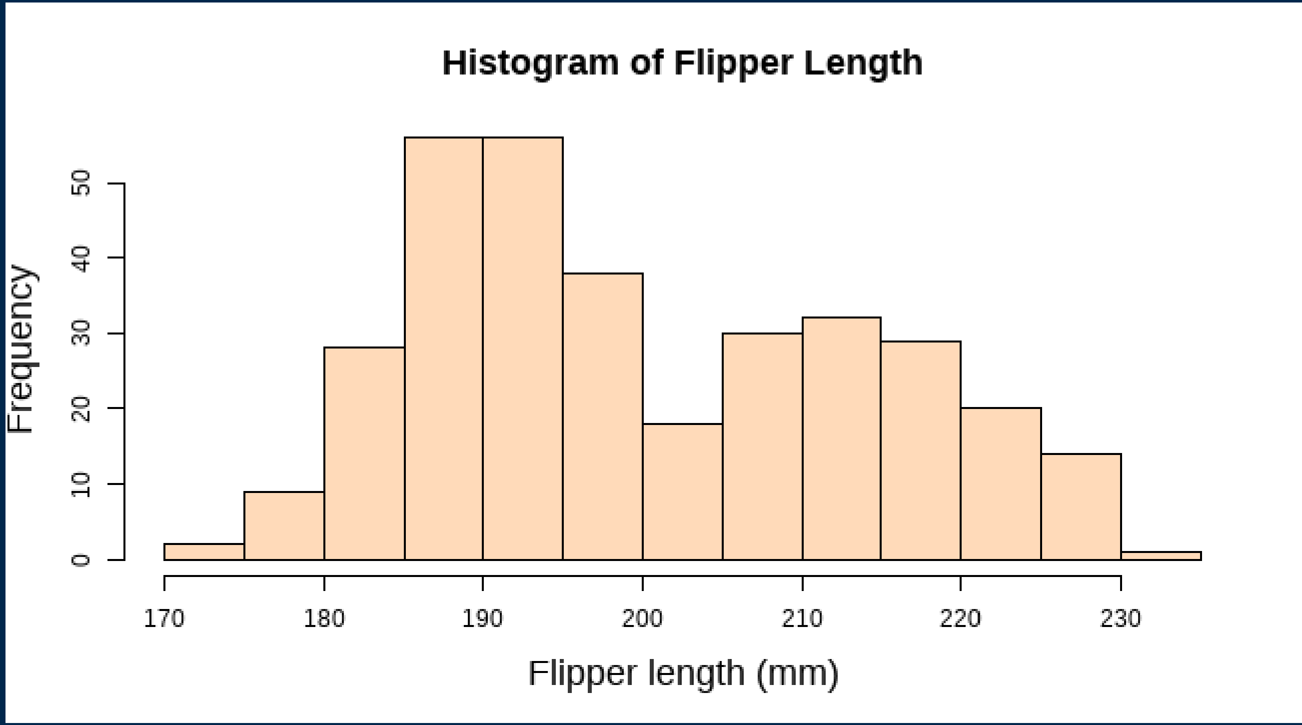
```

	One Sample t-test

data:  textbooks$diff
t = 7.6488, df = 72, p-value = 6.928e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  9.435636 16.087652
sample estimates:
mean of x 
 12.76164 
```

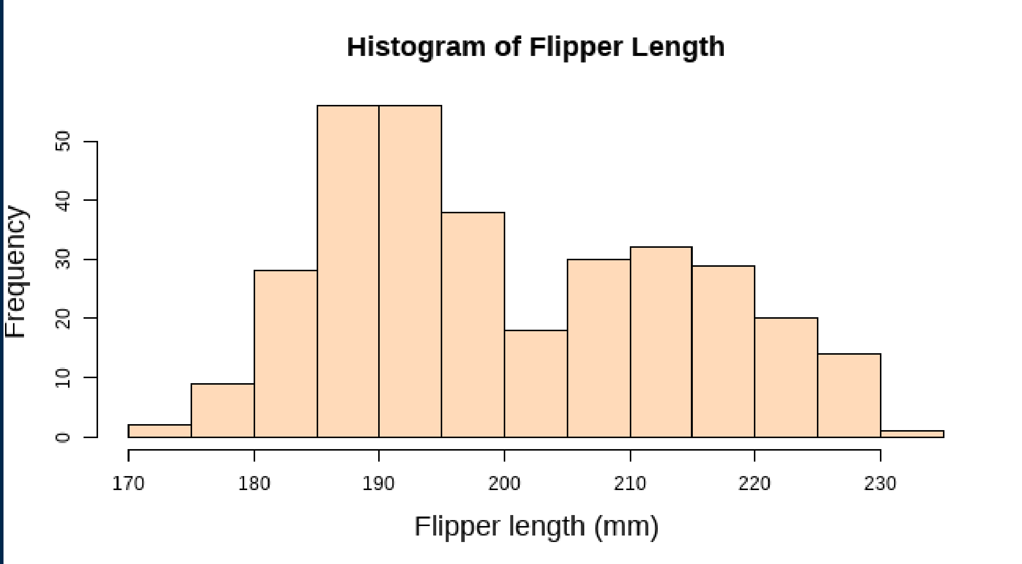
Difference of Two Means: $\mu_1 - \mu_2$

- Read in the penguin data on line ~165
- Remember this bimodal histogram from last week? (line ~171)



Difference of Two Means: $\mu_1 - \mu_2$

```r
hist(penguins$flipper_length_mm,
     main = "Histogram of Flipper Length",
     xlab = "Flipper length (mm)",
     col = "peachpuff")
```

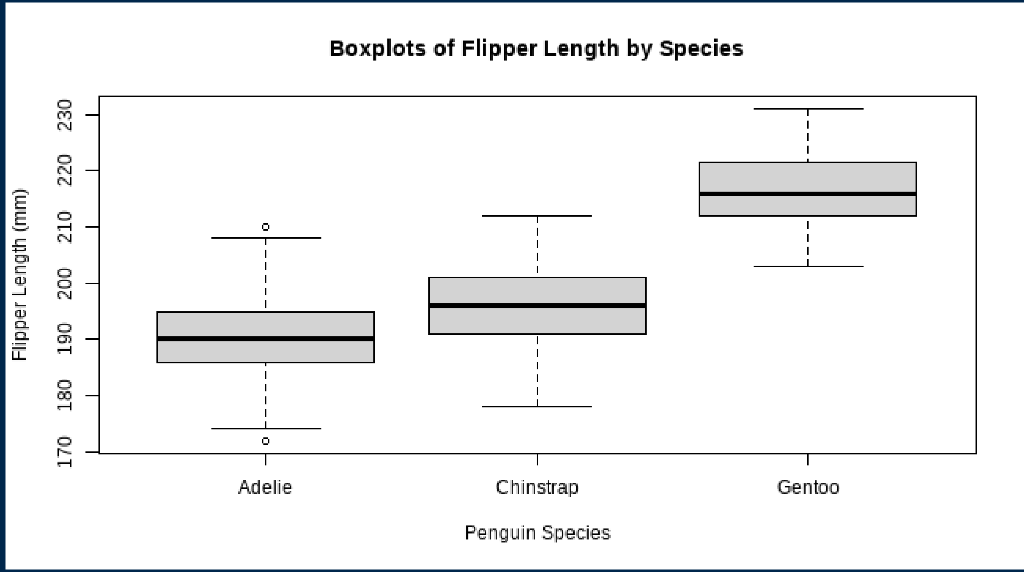


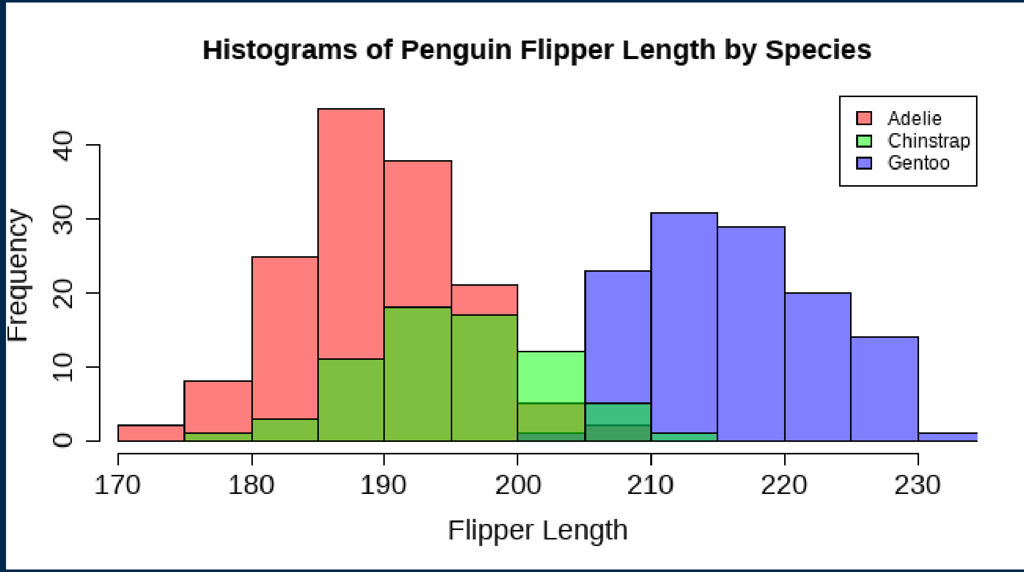
- Bimodal distributions suggest a **subgroup effect**
- There are *three different species* in this data

TASK: Take 2 minutes to write code in the `investigateSpecies` chunk (line 182) to investigate the relationship between species and flipper length.

Difference of Two Means: $\mu_1 - \mu_2$

```r
boxplot(flipper_length_mm ~ species,
        data = penguins,
        xlab = "Penguin Species",
        ylab = "Flipper Length (mm)",
        main = "Boxplots of Flipper Length by
```





(code for this histogram is available on request; it's a little too ugly to show)

Difference of Two Means: $\mu_1 - \mu_2$

Let's just compare mean flipper lengths of Adelie and Chinstrap penguins -- the Gentoos are obviously different, so why bother. **Hypotheses?** (line ~188)

Difference of Two Means: $\mu_1 - \mu_2$

Let's just compare mean flipper lengths of Adelie and Chinstrap penguins -- the Gentoos are obviously different, so why bother. **Hypotheses?** (line ~188)

$$H_0 : \mu_{\text{Adelie}} - \mu_{\text{Chinstrap}} = 0 \quad \text{vs.} \quad H_a : \mu_{\text{Adelie}} - \mu_{\text{Chinstrap}} \neq 0$$

Difference of Two Means: $\mu_1 - \mu_2$

Let's just compare mean flipper lengths of Adelie and Chinstrap penguins -- the Gentoos are obviously different, so why bother. **Hypotheses?** (line ~188)

$$H_0 : \mu_{\text{Adelie}} - \mu_{\text{Chinstrap}} = 0 \quad \text{vs.} \quad H_a : \mu_{\text{Adelie}} - \mu_{\text{Chinstrap}} \neq 0$$

Subset the data to contain just Adelies and Chinstraps (line ~197)

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

```

   Adelie Chinstrap 
      146        68 
```

Difference of Two Means: $\mu_1 - \mu_2$

The most important question in statistics is not whether you **can** do something, it's whether you **should** do it.

Check Conditions!

Difference of Two Means: $\mu_1 - \mu_2$

> The most important question in statistics is not whether you **can** do something, it's whether you **should** do it.

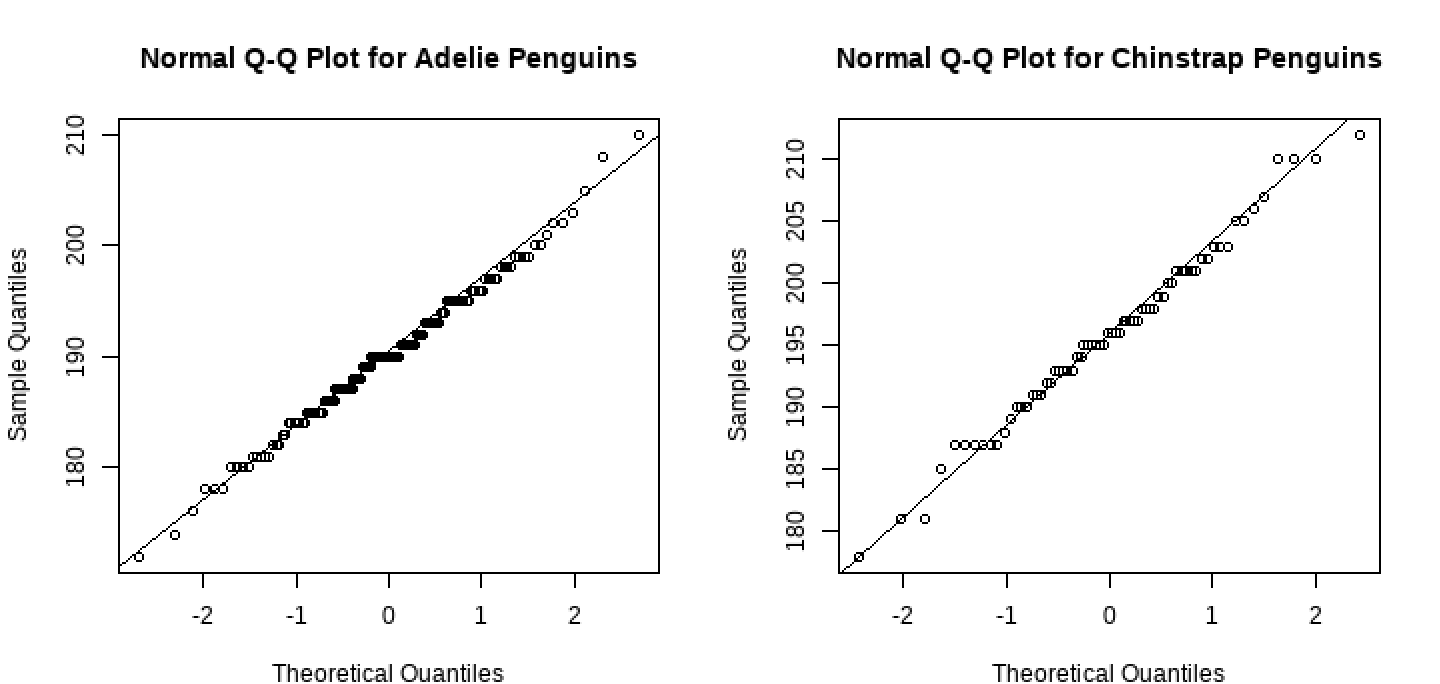
Check Conditions!

1. Independence:
    1. Penguins *within* each species are selected independently
    2. The samples from each species (*between* samples) are independent
2. Nearly Normal:
    1. Adelie flipper lengths are nearly normal
    2. Chinstrap flipper lengths are nearly normal

Difference of Two Means: Check Normality

```
qqnorm(penguinsSubset$flipper_length_mm[penguinsSubset$species == "Adelie"],
       main = "Normal Q-Q Plot for Adelie Penguins")
qqline(penguinsSubset$flipper_length_mm[penguinsSubset$species == "Adelie"])

qqnorm(penguinsSubset$flipper_length_mm[penguinsSubset$species == "Chinstrap"],
       main = "Normal Q-Q Plot for Chinstrap Penguins")
qqline(penguinsSubset$flipper_length_mm[penguinsSubset$species == "Chinstrap"])
```



Two-Sample t-Test

Remember formula notation:

> (response variable) ~ (grouping/explanatory variable)

```
t.test(flipper_length_mm ~ species,
       data = penguinsSubset,
       mu = 0,
       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 is not equal to 0
95 percent confidence interval:
 -7.739129 -3.702450
sample estimates:
   mean in group Adelie mean in group Chinstrap 
               190.1027                195.8235 
```

What's our conclusion?

Code Cheat Sheet 💻

```
pt(q, df, lower.tail = TRUE)
```

- `q` is the x-axis value you want to find an area related to
- `df` is the degrees of freedom of the t distribution
- `lower.tail` determines whether `pt()` finds the area to the left or right of `q`. If `lower.tail = TRUE` (the default), it shades to the left. If `lower.tail = FALSE`, it shades to the right.

Code Cheat Sheet 💻

`qt(q, df, lower.tail = TRUE)`

- p is the probability or area under the curve you want to find an x-axis value for
- `df` is the degrees of freedom of the t distribution
- `lower.tail` determines whether `pt()` finds the area to the left or right of q. If `lower.tail = TRUE` (the default), it shades to the left. If `lower.tail = FALSE`, it shades to the right.

Code Cheat Sheet 💻

`plotT()`

- `df` refers to the degrees of freedom of the distribution to plot. You must provide this value.
- `shadeValues` is a vector of up to 2 numbers that define the region you want to shade
- `direction` can be one of `less`, `greater`, `outside`, or `inside`, and controls the direction of shading between `shadeValues`. Must be `less` or `greater` if `shadeValues` has only one element; `outside` or `inside` if two
- `col.shade` controls the color of the shaded region, defaults to `"cornflowerblue"`
- `...` lets you specify other graphical parameters to control the appearance of the normal curve (e.g., `lwd`, `lty`, `col`, etc.)

Code Cheat Sheet 💻

`qqnorm(y, ...)`

- `y` refers to the variable for which you want to create a Q-Q plot
- `...` lets you control graphical elements of the plot like `pch`, `col`, etc.

`qqline(y, ...)`

- `y` refers to the variable for which you created a Q-Q plot
- `...` lets you control graphical elements of the plot like `pch`, `col`, etc.
- Function can only be used *after* using `qqnorm()`

Code Cheat Sheet 💻

```
t.test(x, alternative = c("two.sided", "less",
"greater"), mu = 0, conf.level = 0.95)
```

- `x` is a vector of data values OR a formula of the form *response ~ group* for two-sample t-tests.
- `alternative` specifies the direction of the alternative hypothesis; must be one of "two.sided", "less", or "greater"
- `mu` indicates the true value of the mean (under the null hypothesis); defaults to 0
- `conf.level` is the confidence level to be used in constructing a confidence interval; must be between 0 and 1, defaults to 0.95

Lab Project ⌨

Your tasks

- Complete the "Try It!" and "Dive Deeper" portions of the lab assignment by copy/pasting and modifying appropriate code from earlier in the document.

How to get help

- Piazza!
- Email your lab instructor (not stats250-miller@umich.edu)

Recap and Reminders 💡

We just learned:

- Name that Scenario is a useful study tool
- How to create a "differences" variable
- How to perform a paired t-test in R
- How to perform a two-sample t-test in R

Task	Due Date	Submission
M-Write 2 Revision	Thursday 11/19 4:59PM ET	Canvas
Lab 12	Friday 11/20 8:00AM ET	Canvas
Homework 9	Friday 11/20 8:00AM ET	course.work

M-Write Office Hours on Canvas!