Lab 8: Paired and two sample t tests

Instructions

- Due date: Friday, Oct 30th 11:59pm. Part 1 of this lab focuses on two data sets sampled from data collected early in the HIV epidemic. Part 2 focuses on conducting a t-test, and compares results from a paired test vs. independent test.

Section I: HIV data

- We have two data sets, both sampled from data collected relatively early in the HIV epidemic.
- Deeks, et al. (1999) performed a longitudinal study of HIV-infected adults undergoing Highly Active Anti-Retroviral Therapy (HAART) at San Francisco General Hospital (SFGH).
- Patients were included in this analysis if they received at least 16 weeks of continuous therapy with an anti-retroviral regimen.
- For both data, the outcome is a measure of severity of the disease, a count of an immune cell type called CD4.

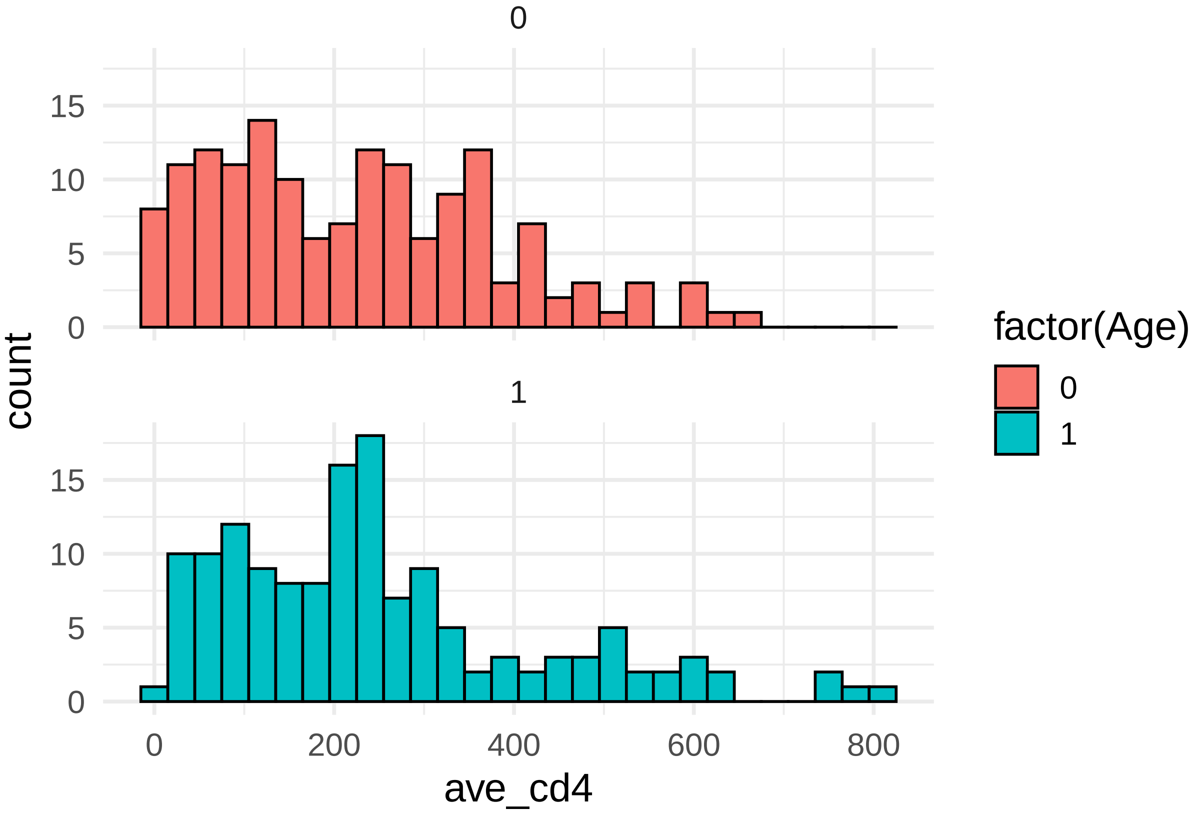
More on data

- The first data set, deeks_ex1.csv, has one response measurement per subject, which is their average CD4 count.
    - The data set also contains a single binary covariate `age` (=1 if $\geq 40 years$, 0 if ≤ 40).
- The second data set, deeks_ex2.csv has two measurements per individual, one at each level of the covariate binary viral load (`vl` $= 1$ if ≥ 2000, `vl` $= 0$ if ≤ 2000).

Age versus CD4 count

1. After importing deeks_ex1.csv into R, compare visually the distribution of CD4 counts between individuals where `age`=1 vs. `age`=0.

```r
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)
library(tidyverse)

deeks <- read.csv("deeks_ex1.csv")
p1 <- ggplot(deeks, aes(x = ave_cd4)) +
  geom_histogram(aes(fill = factor(Age)), binwidth = 30, col = "black") +
  theme_minimal(base_size = 15) + facet_wrap(~Age, nrow = 2)
p1
```



```
check_problem1()
```

```
## [1] "Checkpoint 1 Passed: Correct"
##
## Problem 1
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

2. Which of the testing procedures that we've learned so far can be used to test the difference between the mean CD4 counts across individuals with `age=1` vs. `age=0`? Perform the test using an R testing function. Note the estimated mean difference and the provided 95% confidence interval. Report your p-value rounded to 2 decimal places.

(If you have extra time, confirm that you can calculate the test statistic using dplyr functions only.)

```
# YOUR T-TEST CODE HERE

pvalue_deeks <- 0.21

#H0 = mean_age0 = mean_age1, H1 = mean_age0 not = mean_age1

t.test(deeks %>% filter(Age == 0) %>% pull(ave_cd4),
       deeks %>% filter(Age == 1) %>% pull(ave_cd4),
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  deeks %>% filter(Age == 0) %>% pull(ave_cd4) and deeks %>% filter(Age == 1) %>% pull(ave_cd4)
## t = -1.2563, df = 286.52, p-value = 0.21
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -62.21788  13.73708
## sample estimates:
## mean of x mean of y
##  225.9020  250.1424
```

```
check_problem2()
```

```
## [1] "Checkpoint 1 Passed: Correct"
##
## Problem 2
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

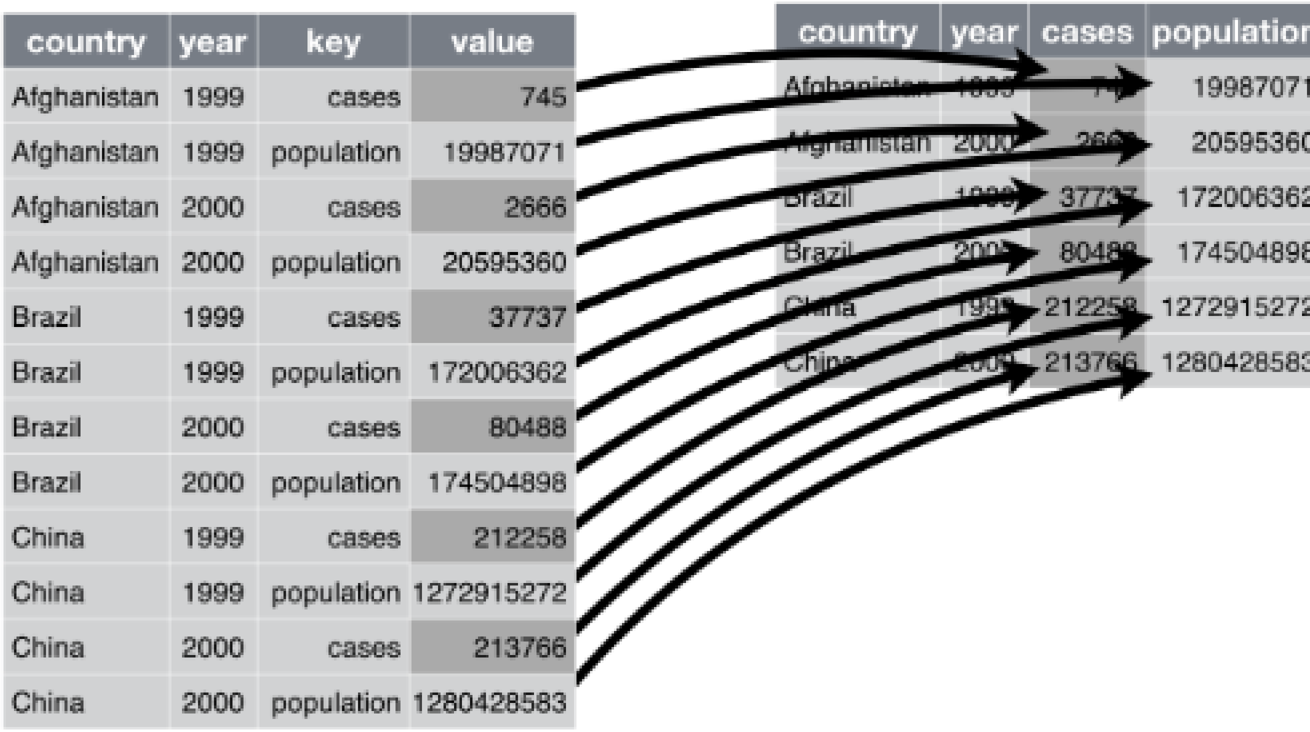
From the t.test output, we estimate the mean difference from the sample to be 225.9020 - 250.1424 = -24.24 with a 95% CI -62.22 to 13.74. Assuming that the null hypothesis of no difference between the groups' CD4 counts is true, there is a 21% chance of seeing a difference between the sample mean of this magnitude or larger.Thus, there is not much evidence against the null hypothesis.

CD4 count and viral load

3.1 After reading in deeks_ex2.csv, visualize the distribution of *individual differences* in CD4 counts during periods of high vs. low viral load measurement. To do this, first note that the data is in long format (with two rows per individual, one for each level of `vl`). To calculate the difference in CD4 count for each individual across levels of `vl` we need to convert the data into "wide" format so that the CD4 measures at `vl=0` and `vl=1` are contained in the same row for each individual. To do this, you need to use the `spread()` function from `tidyr`. Your GSI will help with this if you can't figure it out!

Here is an illustration of how spread works:

```
knitr::include_graphics("lab08-spread-function.png")
```



country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

country	year	cases	populatio
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

table2

```r
## PUT YOUR CODE HERE
dat2_long <- read.csv("deeks_ex2.csv")

dat2_wide <- dat2_long %>% spread(medvl, cd4)
names(dat2_wide) <- c("id", "cd4_LowVL", "cd4_HighVL")
dat2_wide <- mutate(dat2_wide, diff = cd4_HighVL - cd4_LowVL)

# This question is not autograded.
```

4. Which of the testing procedures that we've learned so far can be used to test the difference between each individual's CD4 count during a time of high vs. low viral load? Perform the test using an R testing function. Note the estimated mean difference and the provided 95% confidence interval. Report your p-value rounded to 4 decimal places.

```r
## PUT YOUR T-TEST CODE HERE

t.test(dat2_wide %>% pull(cd4_HighVL), dat2_wide %>% pull(cd4_LowVL),
  alternative = "two.sided", paired = T)
```

```
##
##  Paired t-test
##
```

```
## data:  dat2_wide %>% pull(cd4_HighVL) and dat2_wide %>% pull(cd4_LowVL)
## t = -3.0391, df = 70, p-value = 0.003335
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -79.96732 -16.59606
## sample estimates:
## mean of the differences
##               -48.28169
```

```
pvalue_dat2 <- 0.0033

check_problem4()
```

```
## [1] "Checkpoint 1 Passed: Correct"
##
## Problem 4
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

The paired t-test has a t-statistic of -3.0391for a p-value of 0.0033. Assuming the null hypothesis is true (i.e., no difference between the mean individual difference at low vs high viral load), there is only 0.3% chance of seeing the estimated mean difference -48.28169 that we saw in these samples (or a result more extreme). Given the chance is low, there is strong evidence in favor of the alternative hypothesis that the CD4 count is different between low and high levels of viral load. The estimated difference is -48.28169 with a 95% CI of -79.96732 to -16.59606, supporting the hypothesis of a drop in CD4 count for subjects reporting as viral load changes from low to high.

Section II: Coin Flip Game

Go to this website

The game: See how many dots you can hit in the grid within 30 seconds. We will each try this once with our dominant hand and once with our non-dominant hand (**where your dominant hand is the one you prefer to operate a computer mouse or track pad with**).

Instructions:

Flip a coin to see which hand to play the game with first: - Heads = dominant hand first - Tails = non-dominant hand first

2. Push the **Start Game** button. It will start a timer counting down from 30 seconds. During that time use only the specified hand to click the moving dot as fast as you can. After 30 seconds, the game will stop and display the number of dots that you hit. Record that number in the shared google sheet. **Make sure you put it in the correct column!**. Also fill out the last column of the dataset "Dominant_hand_first". Set this variable to TRUE if you used your dominant hand in the first game or FALSE if you used your non-dominant hand in the first game.

3. Re-do the game, this time with the other hand. Record the results in the spreadsheet.

4. Read the data from the google sheet into R.

Lab 101B: https://docs.google.com/spreadsheets/d/1IxybE5KAHHwLKNni5 edit?usp=sharing

Lab 102B: https://docs.google.com/spreadsheets/d/1Ao2Y9sSwGlguHDct2I4 edit?usp=sharing

Lab 103B: https://docs.google.com/spreadsheets/d/1OmUtQZ79Dx68Gfl8kJ edit?usp=sharing

Lab 104B: https://docs.google.com/spreadsheets/d/1ZxsNxSLv514xfHyK3N lb0dyy8Q4vAjGnH6zZU/edit?usp=sharing

Lab 105B: https://docs.google.com/spreadsheets/d/1qAeUPN6PsvVHPgRW rCEWSRm1blUbzPDAqaVn0B0gNI/edit?usp=sharing

Lab 106B: https://docs.google.com/spreadsheets/d/1rgY7CEtvRUSvVD6mV edit?usp=sharing

Lab 107B: https://docs.google.com/spreadsheets/d/1z8onu78ZNzv_RlwyYPsnrch8lidtlABvQDnCg0I9jeQ/edit?usp=sharing

Lab 108B: https://docs.google.com/spreadsheets/d/1L2e1X7BQBvK5QgjFie Oncv0/edit?usp=sharing

Lab 109B: https://docs.google.com/spreadsheets/d/1dQes48BgRpt9FjLOeLI mLv0jaiOmBV-fsk/edit?usp=sharing

Lab 110B: https://docs.google.com/spreadsheets/d/1pjBrYQG6ObIRmcaRP dEpogE/edit?usp=sharing

sample: https://docs.google.com/spreadsheets/d/1v9Mvm2hAOB3orINrcbV edit?usp=sharing

```r
library(googlesheets)
library(dplyr)
```

```r
#Lab101B: my_key <- "1IxybE5KAHHwLKNni5jFErT8e7JXp1nhYB64kHL6Sheg"
#Lab102B: my_key <- "1Ao2Y9sSwGlguHDct2I4z4c6q6TPETTYTsfP7Uy6o5L0"
```

```
#Lab103B: my_key <- "1OmUtQZ79Dx68Gfl8kJgAZrfPN6mAnveDTAIHBUY6tWw"
#Lab104B: my_key <- "1ZxsNxSLv514xfHyK3NYtBF5_lb0dyy8Q4vAjGnH6zZU"
#Lab105B: my_key <- "1qAeUPN6PsvVHPgRW-rCEWSRm1blUbzPDAqaVn0B0gNI"
#Lab106B: my_key <- "1rgY7CEtvRUSvVD6mVFRi003u4Uba0fae2mJammCfdEg"
#Lab107B: my_key <- "1z8onu78ZNzv_RlwyYPsnrch8lidtlABvQDnCg0I9jeQ"
#Lab108B: my_key <- "1L2e1X7BQBvK5QgjFieZeiUX0mq6D9ly8eDpVW-Oncv0"
#Lab109B:
  my_key <- "1dQes48BgRpt9FjL0eLHh2OedqpR_mLv0jai0mBV-fsk"
#Lab110B: my_key <- "1pjBrYQG60bIRmcaRPnualvpv8isf02N7uoa6-dEpogE"
```

##Remove eval = F from the chunk header before moving on!

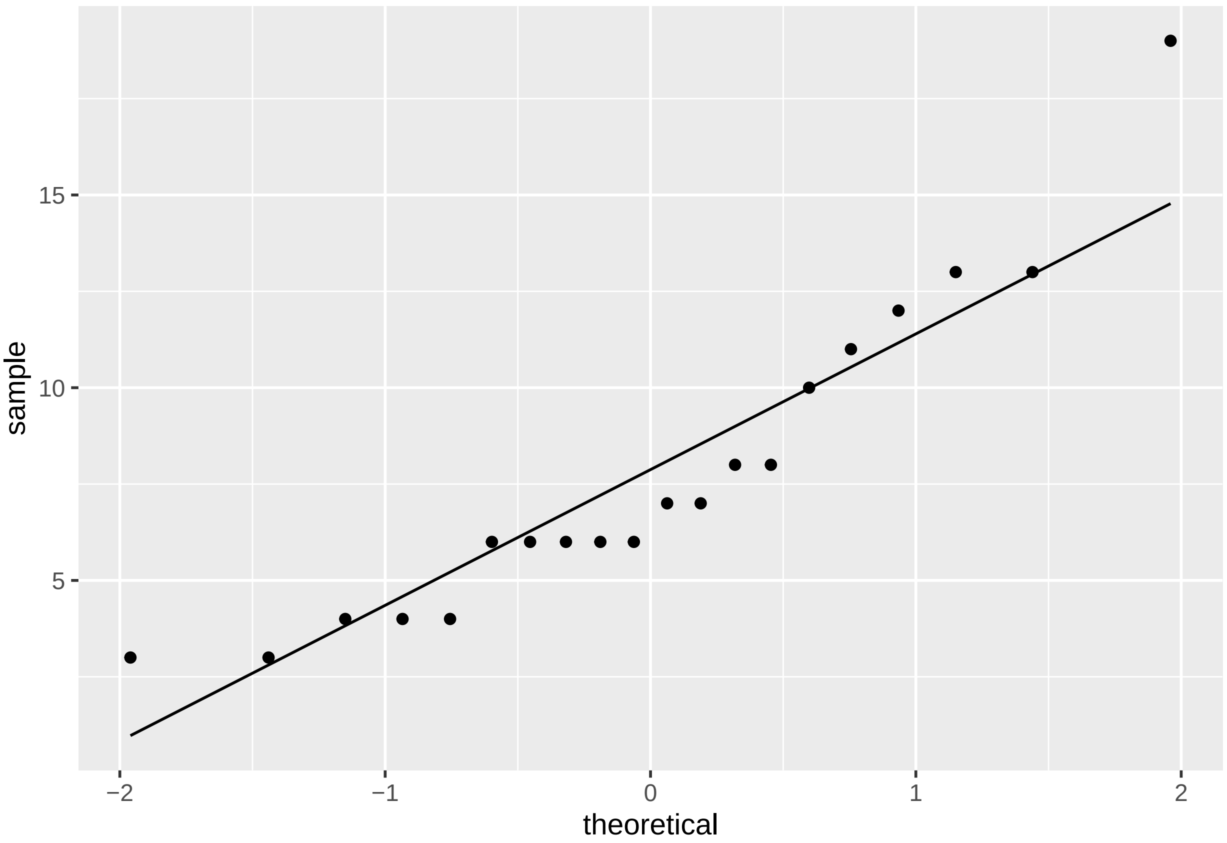
```
our_sheet <- my_key %>%
  gs_key(lookup = FALSE) %>%
  gs_read(range = "A1:D100")
```

```
## Worksheets feed constructed with public visibility

## Warning: ‘as_data_frame()‘ is deprecated as of tibble 2.0.0.
## Please use ‘as_tibble()‘ instead.
## The signature and semantics have changed, see ‘?as_tibble‘.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.

## Accessing worksheet titled ’Sheet1’.

## Warning: ‘select_()‘ is deprecated as of dplyr 0.7.0.
## Please use ‘select()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.

## Warning: ‘arrange_()‘ is deprecated as of dplyr 0.7.0.
## Please use ‘arrange()‘ instead.
## See vignette(’programming’) for more help
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.

## Warning: ‘filter_()‘ is deprecated as of dplyr 0.7.0.
## Please use ‘filter()‘ instead.
## See vignette(’programming’) for more help
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.

##
## -- Column specification --------------------------------------------------------
## cols(
##   Student_name = col_character(),
##   Dominant_num_dots_hit = col_double(),
##   Non_dominant_num_dots_hit = col_double(),
##   Dominant_hand_first = col_logical()
## )
```

5. These data are very naturally paired. What two assumptions do we need to make to use a paired t-test? For each assumption, either write why you think the assumption is met (or not met), or investigate the assumption by creating a plot, and comment on whether the plot supports the assumption.

1) It is a SRS.
2) The differencnes in the population are Normally distributed.

```
## PUT YOUR CODE HERE: Write your code here to investigate the other assumption.
## Hint: You need to first compute a new variable using dplyr before you make your plot :).

our_sheet  <- mutate(our_sheet, diff = Dominant_num_dots_hit - Non_dominant_num_dots_hit)
ggplot(our_sheet, aes(sample = diff)) + geom_qq() + geom_qq_line()
```



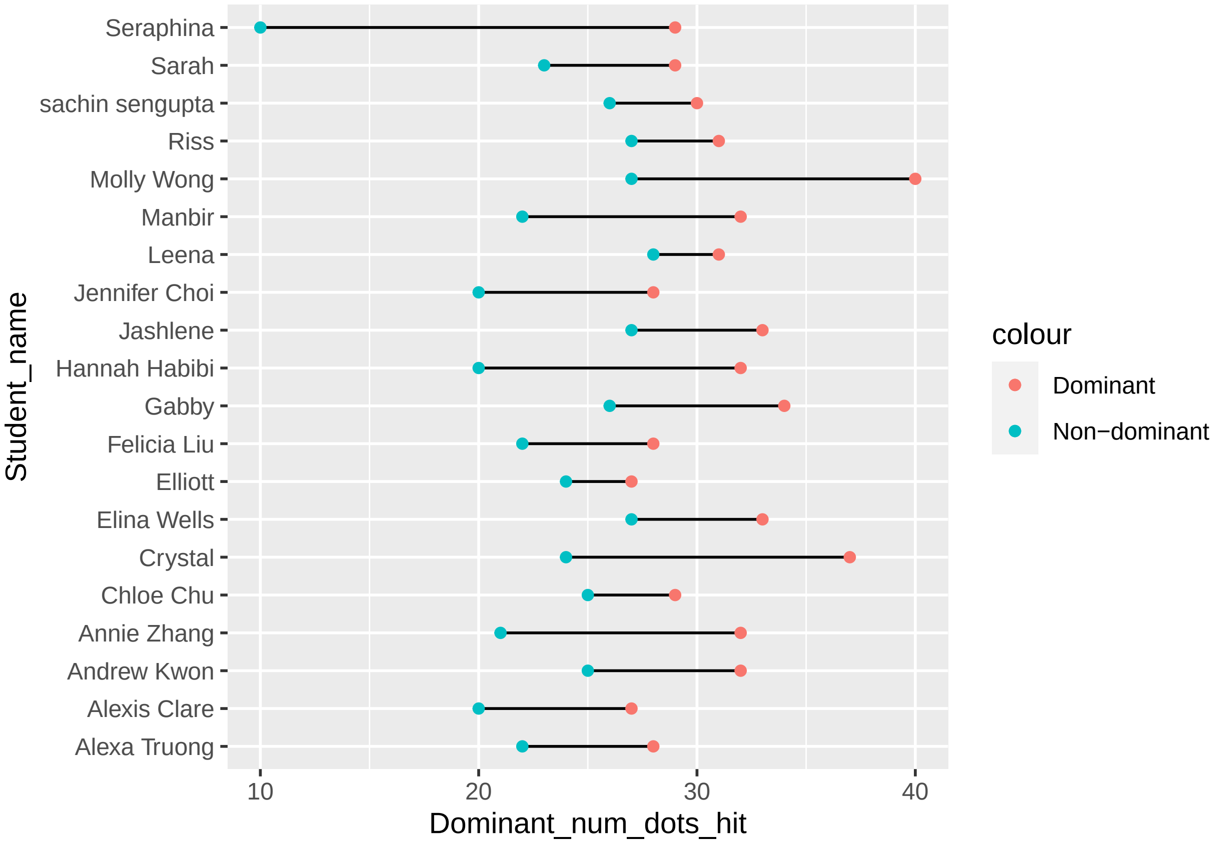
1) The students in the current lab are randomly selected from the 300 students in our class, so the sample is a SRS.
2) According to the ggplot, the individual differences are Normally distributed.

6. Before performing the test, take a look at the data by making a "dumbbell" plot. This type of plot has student name on the y-axis, and the number of dots hit on the x axis. For each student you put a point at the two reaction times and connect them with a line. Here is the code to make the plot. We can also color the points by hand dominance. Based on the plot, comment on whether there appears to be a significant difference between the number of points hit between the dominant and non-dominant hand.

Here is the code to make the dumbbell chart. You will need to change `our_sheet` to the name of your saved dataset (if you changed the name).

##Remove eval = F from the chunk header before moving on!

```
# This code is provided to students because it is a bit advanced.
# You are not expected to know how to make this plot yourself!


ggplot(data = our_sheet, aes(x = Dominant_num_dots_hit, y = Student_name)) +
  geom_segment(aes(xend = Non_dominant_num_dots_hit, yend = Student_name)) +
  geom_point(aes(col = "Dominant")) +
  geom_point(aes(x = Non_dominant_num_dots_hit, col = "Non-dominant"))
```



For these data, all of us had a higher number of dots for our dominant hand vs our non-dominant hand. We expect the tests would reject the null hypothesis of no differences between the two hands.

7. Use R to conduct a paired two-sided t-test on the data, and note the 95% confidence interval for the test. Report your p-value rounded to 2 decimal places. Interpret the p-value and the confidence interval for the test.

```
## PUT YOUR T-TEST CODE HERE

t.test(our_sheet %>% pull(Dominant_num_dots_hit), our_sheet %>%
      pull(Non_dominant_num_dots_hit),
       alternative = "two.sided", paired = T)
```

```
##
##  Paired t-test
```

```
##
## data:  our_sheet %>% pull(Dominant_num_dots_hit) and our_sheet %>% pull(Non_dominant_num_dots_hit)
## t = 8.5105, df = 19, p-value = 6.608e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.881712 9.718288
## sample estimates:
## mean of the differences
##                     7.8
```

```
pvalue_paired <- 0.00


check_problem7()
```

```
## [1] "Checkpoint 1 Passed: Correct"
##
## Problem 7
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Our p-value for the test is less than 0.0001. This means that under the null hypothesis of no difference between the hands in the number of dots hit, there is a less than 0.0001% chance of observing the difference (or more greater magnitude) that we saw between the sample average of individual differences. The probability is so small that we reject the null hypothesisin favor of the alternative hypothesis that there is a difference between hands in the number of dots hit.

The 95% CI for the differences is 5.881712 to 9.718288. If our model assumptions are correct (SRS & Normally distributed) and there is only random error affecting the estimate, the method of calculating confidence intervals will contain the true underlying value of difference between dominant and non-dominant hands 95% of the time.

8. Re-run the code for the test, but this time set `paired=F`, which is incorrect. The reason we want to run the incorrect test is to compare the p-value from this test to the p-value from the paired t-test. Is it smaller or larger? Why is that?

```
# YOUR T-TEST CODE HERE

t.test(our_sheet %>% pull(Dominant_num_dots_hit), our_sheet %>%
       pull(Non_dominant_num_dots_hit),
       alternative = "two.sided", paired = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  our_sheet %>% pull(Dominant_num_dots_hit) and our_sheet %>% pull(Non_dominant_num_dots_hit)
## t = 6.6171, df = 36.463, p-value = 9.879e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.410419 10.189581
## sample estimates:
## mean of x mean of y
##      31.1      23.3
```

```
# Then, uncomment one of these choices:
# p8 <- "smaller"
p8 <- "larger"


check_problem8()
```

```
## [1] "Checkpoint 1 Passed: Correct"
##
## Problem 8
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

9. Lastly, we didn't use the data on the last column in the data frame, which recorded whether you were randomized to using your dominant hand first. Why might this matter? What could we have done to investigate whether it mattered?

This might matter because whether we use our dominant hand or non-dominant hand could make a difference in the number of dots that we hit. For example, we could become better at hitting the dots after doing it for the first time with one hand, and that could affect the results of our other hand. We could have ran tests for those who used our dominant hand first and ran another test for those who used our non-dominant hand first to see if there is a difference in the statistics and CI's that we calculate.

Check your score

Click on the middle icon on the top right of this code chunk (with the downwards gray arrow and green bar) to run all your code in order. Then, run this chunk to check your score.

```
# Just run this chunk.
total_score()
```

```
##                      Test Points_Possible          Type
## Problem 1          PASSED               1    autograded
## Problem 2          PASSED               1    autograded
## Problem 3 NOT YET GRADED               1 free-response
## Problem 4          PASSED               1    autograded
## Problem 5 NOT YET GRADED               1 free-response
## Problem 6 NOT YET GRADED               1 free-response
## Problem 7          PASSED               1    autograded
## Problem 8          PASSED               1    autograded
## Problem 9 NOT YET GRADED               1 free-response
```

Submission

For assignments in this class, you'll be submitting using the **Terminal** tab in the pane below. In order for the submission to work properly, make sure that:

1. Any image files you add that are needed to knit the file are in the `src` folder and file paths are specified accordingly.
2. You **have not changed the file name** of the assignment.
3. The file is saved (the file name in the tab should be **black**, not red with an asterisk).
4. The file knits properly.

Once you have checked these items, you can proceed to submit your assignment.

1. Click on the **Terminal** tab in the pane below.
2. Copy-paste the following line of code into the terminal and press enter.

```
cd; cd ph142-fa20/lab/lab08; python3 turn_in.py
```

3. Follow the prompts to enter your Gradescope username and password. When entering your password, you won't see anything come up on the screen–don't worry! This is just for security purposes–just keep typing and hit enter.
4. If the submission is successful, you should see "Submission successful!" appear as output.
5. If the submission fails, try to diagnose the issue using the error messages–if you have problems, post on Piazza.

The late policy will be strictly enforced, **no matter the reason**, including submission issues, so be sure to submit early enough to have time to diagnose issues if problems arise.