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.

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

deeks <- "LOAD DATA HERE"
p1 <- "YOUR GGPLOT CODE HERE"
p1
```

```
## [1] "YOUR GGPLOT CODE HERE"
```

```
check_problem1()
```

```
## [1] "Checkpoint 1 Error: Assign your ggplot to p1!"
##
## Problem 1
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

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 <- "REPLACE WITH P-VALUE ROUNDED TO 2 DECIMAL PLACES"


check_problem2()
```

```
## [1] "Checkpoint 1 Error: Incorrect"
##
## Problem 2
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

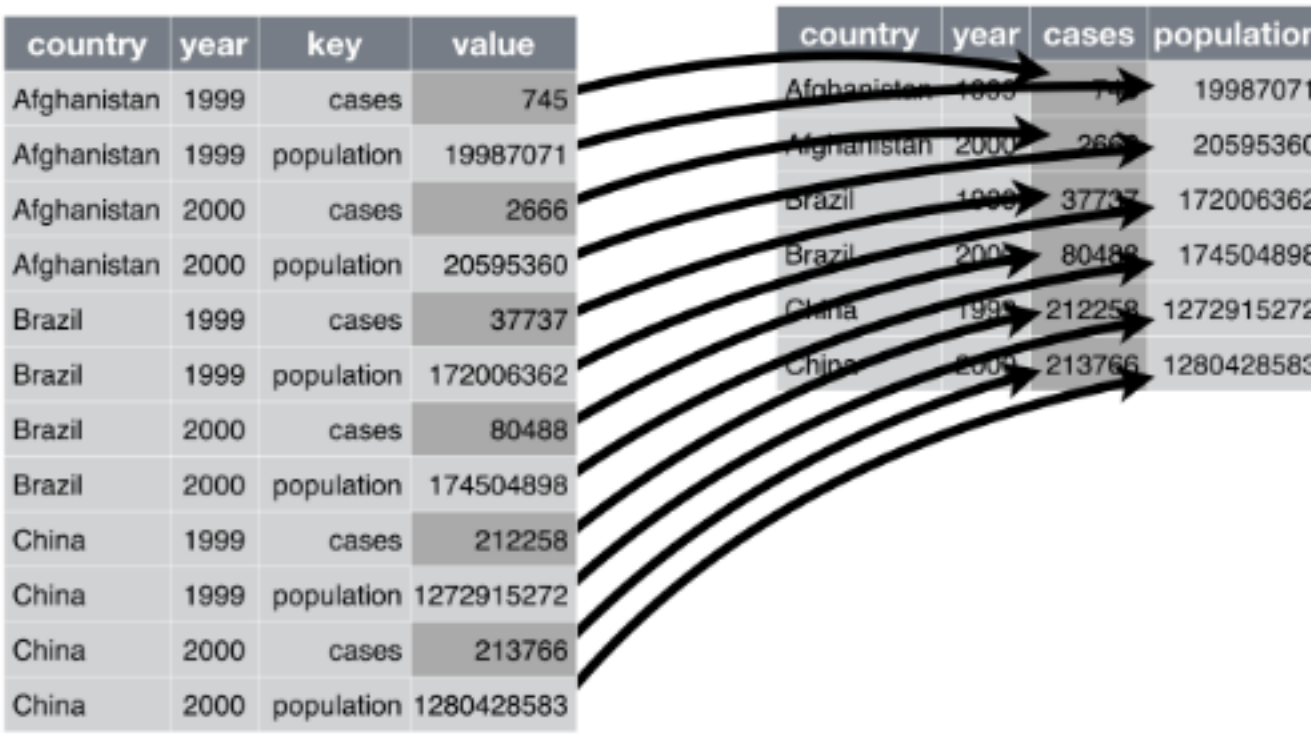
[TODO: YOUR ANSWER HERE]

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

table2


```
## PUT YOUR CODE HERE


# 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.

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

pvalue_dat2 <- "REPLACE WITH P-VALUE ROUNDED TO 4 DECIMAL PLACES"

check_problem4()
```

```
## [1] "Checkpoint 1 Error: Incorrect"
##
## Problem 4
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
```

```
## --------
## Test: FAILED
```

[TODO: YOUR ANSWER HERE]

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/1dQes48BgRpt9FjLOeLH
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

```
library(googlesheets)
```

```
## Warning: package 'googlesheets' was built under R version 4.0.2
```

```
library(dplyr)
```

```
#Lab101B: my_key <- "1IxybE5KAHHwLKNni5jFErT8e7JXp1nhYB64kHL6Sheg"
#Lab102B: my_key <- "1Ao2Y9sSwGlguHDct2I4z4c6q6TPETTYTsfP7Uy6o5L0"
#Lab103B: my_key <- "1OmUtQZ79Dx68Gfl8kJgAZrfPN6mAnveDTAIHBUY6tWw"
#Lab104B: my_key <- "1ZxsNxSLv514xfHyK3NYtBF5_lbOdyy8Q4vAjGnH6zZU"
#Lab105B: my_key <- "1qAeUPN6PsvVHPgRW-rCEWSRm1blUbzPDAqaVnOBOgNI"
#Lab106B: my_key <- "1rgY7CEtvRUSvVD6mVFRiOO3u4UbaOfae2mJammCfdEg"
#Lab107B: my_key <- "1z8onu78ZNzv_RlwyYPsnrch8lidtlABvQDnCg0I9jeQ"
#Lab108B: my_key <- "1L2e1X7BQBvK5QgjFieZeiUXOmq6D9ly8eDpVW-OncvO"
#Lab109B: my_key <- "1dQes48BgRpt9FjLOeLHh2OedqpR_mLvOjaiOmBV-fsk"
#Lab110B: my_key <- "1pjBrYQG6ObIRmcaRPnualvpv8isfO2N7uoa6-dEpogE"
```

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

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

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.

[LIST ASSUMPTIONS HERE]

```
## 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 :).
```

[Comment on assumptions met/unmet here]

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"))
```

[TODO: YOUR ANSWER HERE]

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

pvalue_paired <- "REPLACE WITH YOUR PVALUE ROUNDED TO 2 DECIMAL PLACES"


check_problem7()
```

```
## [1] "Checkpoint 1 Error: Incorrect"
##
## Problem 7
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

[TODO: YOUR ANSWER HERE]

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

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


check_problem8()
```

```
## [1] "Checkpoint 1 Error: Incorrect"
##
## Problem 8
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

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?

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         FAILED               1    autograded
## Problem 2         FAILED               1    autograded
## Problem 3 NOT YET GRADED               1 free-response
## Problem 4         FAILED               1    autograded
## Problem 5 NOT YET GRADED               1 free-response
## Problem 6 NOT YET GRADED               1 free-response
## Problem 7         FAILED               1    autograded
## Problem 8         FAILED               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.