

Lab 8: T-Test

```
BEGIN ASSIGNMENT
requirements: requirements.R
generate: true
files:
- data
- turn_in.py
- src
```

Instructions

- Due date: Tuesday, April 6 at 10:00pm PST with 2 hour grace period.
- Late penalty: 50% late penalty if submitted within 24 hours of due date, no marks for assignments submitted thereafter.
- This assignment is graded on **correct completion**, all or nothing. You must pass all public tests and submit the assignment for credit.
- Submission process: Follow the submission instructions on the final page. Make sure you do not remove any `\newpage` tags or rename this file, as this will break the submission.

Introduction

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 ≥ 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 (`v1` = 1 if ≥ 2000 , `v1` = 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`. Note that these datasets are located in the data folder.

BEGIN QUESTION

name: p1

manual: true

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

```
. = " # BEGIN PROMPT
deeks1 <- 'LOAD DATA HERE'
p1 <- 'YOUR GG PLOT CODE HERE'
p1
" # END PROMPT
```

```
# BEGIN SOLUTION NO PROMPT
```

```
deeks1 <- read_csv("data/deeks_ex1.csv")
```

```
## Parsed with column specification:
```

```
## cols(
```

```
##   id = col_double(),
```

```
##   ave_cd4 = col_double(),
```

```
##   Age = col_double()
```

```
## )
```

```
p1 <- ggplot(deeks1, aes(x = ave_cd4)) +
  geom_histogram(aes(fill = factor(Age)), binwidth = 30, col = "black") +
  theme_minimal(base_size = 15) +
  facet_wrap(~factor(Age), nrow = 2)
```

```
# p1 <- ggplot(deeks1, aes(y = ave_cd4)) +
#   geom_boxplot(aes(fill = factor(Age)), col = "black") +
#   theme_minimal(base_size = 15)
# END SOLUTION
```

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.)

BEGIN QUESTION

name: p2

manual: false

points: 1

```

. = " # BEGIN PROMPT
# YOUR T-TEST CODE HERE

pvalue_deeks <- 'REPLACE WITH P-VALUE ROUNDED TO 2 DECIMAL PLACES'
" # END PROMPT

# BEGIN SOLUTION NO SOLUTION
t.test(deeks1 %>% filter(Age == 0) %>% pull(ave_cd4),
       deeks1 %>% filter(Age == 1) %>% pull(ave_cd4),
       alternative = "two.sided")

##
## Welch Two Sample t-test
##
## data: deeks1 %>% filter(Age == 0) %>% pull(ave_cd4) and deeks1 %>% 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

pvalue_deeks <- 0.21
# END SOLUTION

## Test ##
test_that("p2a", {
  expect_true(pvalue_deeks > 0 & pvalue_deeks < 1)
  print("Checking: range of pvalue_deeks")
})

## [1] "Checking: range of pvalue_deeks"
## Test passed

## Test ##
test_that("p2b", {
  expect_true(all.equal(pvalue_deeks, 0.21, 0.001))
  print("Checking: value of pvalue_deeks")
})

## [1] "Checking: value of pvalue_deeks"
## Test passed

```

From the `t.test` output you can estimate the mean difference from the sample to be $225.9020 - 250.1425 = -24.24$ 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 means 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 `v1`). To calculate the difference in CD4 count for each individual across levels of `v1` we need to convert the data into “wide” format so that the CD4 measures at `v1=0` and `v1=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!

BEGIN QUESTION

name: p3

manual: true

Here is an illustration of how `spread` works:

```
knitr::include_graphics("src/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	population
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

```
# BEGIN SOLUTION
deeks2_long <- read_csv("data/deeks_ex2.csv")
```

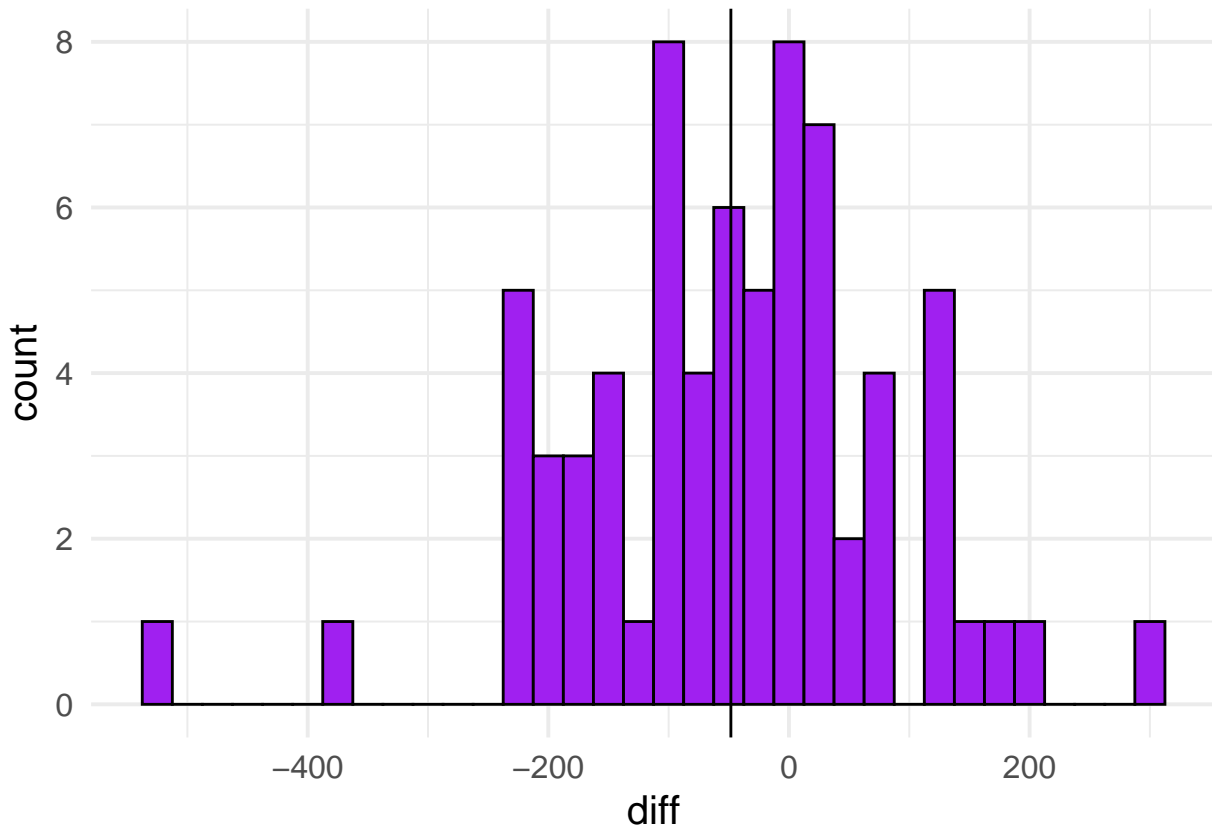
```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   cd4 = col_double(),
##   medv1 = col_double()
## )
```

```

deeks2_wide <- deeks2_long %>% spread(medvl, cd4)
names(deeks2_wide) <- c("id", "cd4_LowVL", "cd4_HighVL")
deeks2_wide <- mutate(deeks2_wide, diff = cd4_HighVL - cd4_LowVL)

# all differences
ggplot(deeks2_wide, aes(x = diff)) +
  geom_histogram(binwidth = 25, col = "black", fill = "purple") +
  theme_minimal(base_size = 15) +
  geom_vline(xintercept = mean(deeks2_wide$diff))

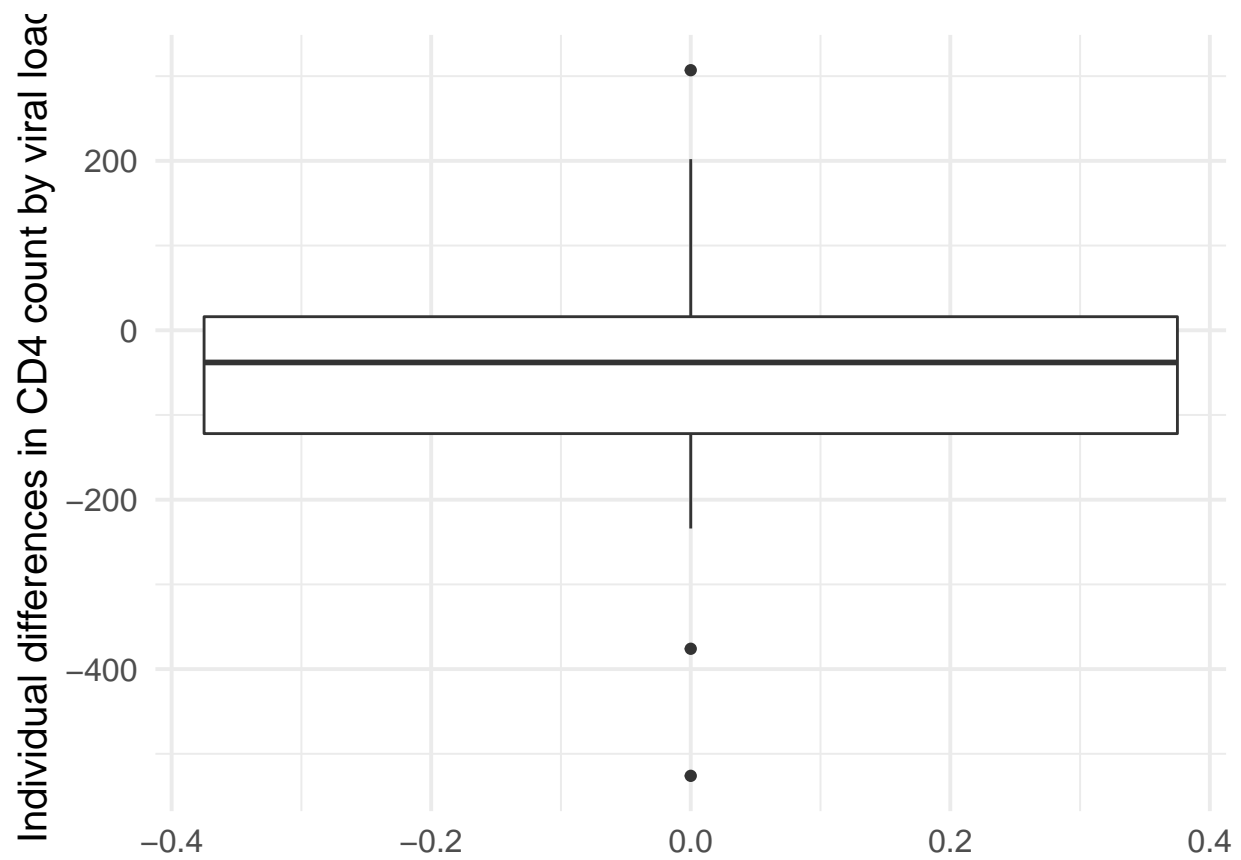
```



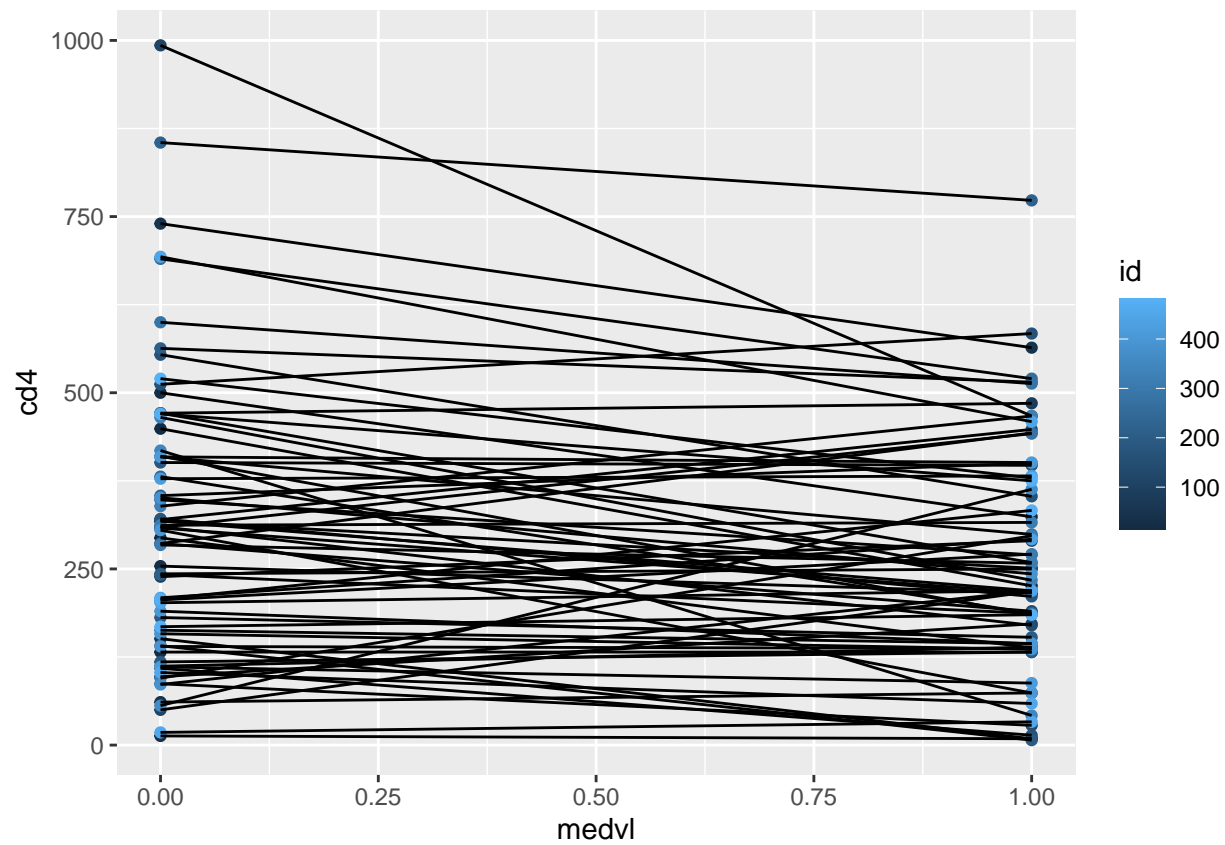
```

ggplot(deeks2_wide, aes(y = diff)) + geom_boxplot() +
  theme_minimal(base_size = 15) +
  labs(y = "Individual differences in CD4 count by viral load")

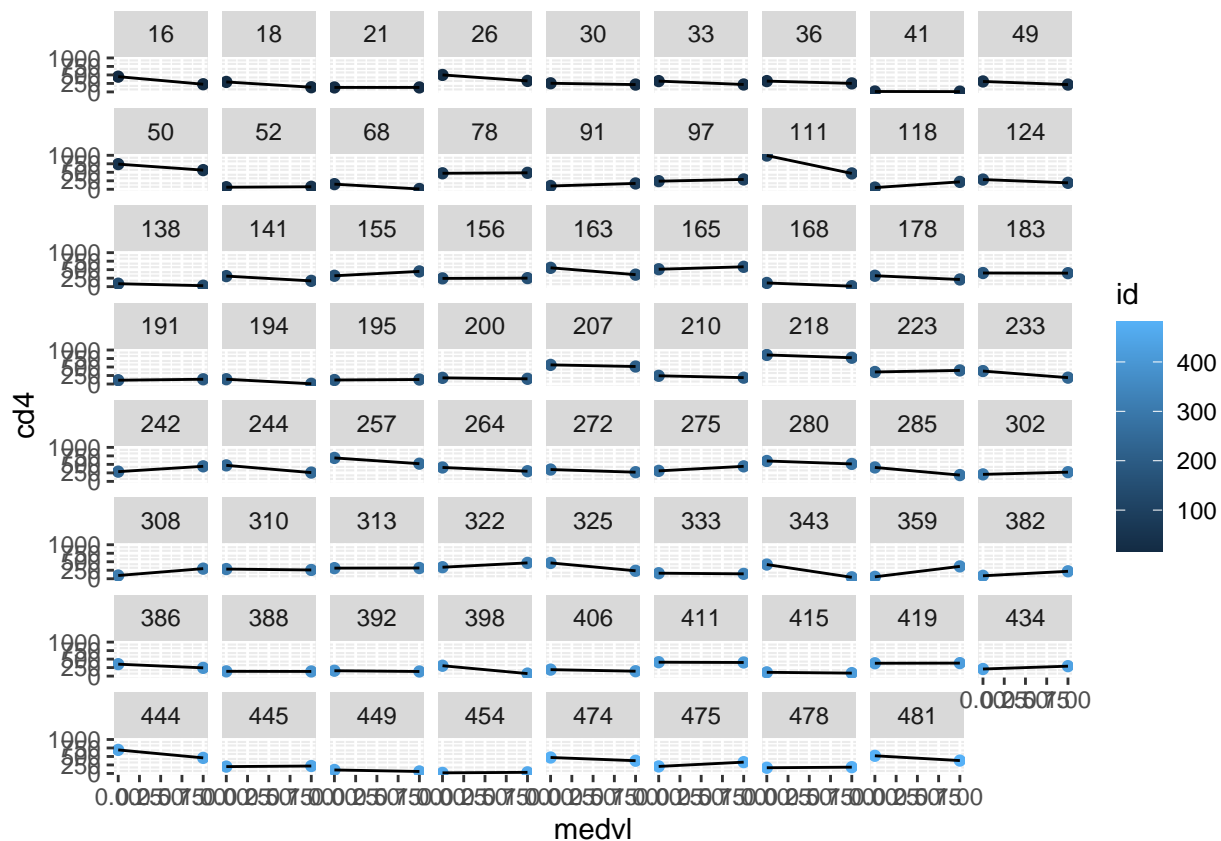
```



```
# individual differences  
# spaghetti plots  
ggplot(deeks2_long, aes(y = cd4, x = medvl)) + geom_point(aes(col = id)) +  
  geom_line(aes(group = id))
```



```
ggplot(deeks2_long, aes(y = cd4, x = medvl)) + geom_point(aes(col = id)) +  
  geom_line(aes(group = id)) + facet_wrap(~id)
```



END SOLUTION

- 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.

BEGIN QUESTION

name: p4

manual: false

points: 1

```
. = " # BEGIN PROMPT
## PUT YOUR T-TEST CODE HERE

pvalue_deeks2 <- 'REPLACE WITH P-VALUE ROUNDED TO 4 DECIMAL PLACES'
" # END PROMPT

# BEGIN SOLUTION NO PROMPT
t.test(deeks2_wide %>% pull(cd4_HighVL),
       deeks2_wide %>% pull(cd4_LowVL),
       alternative = "two.sided",
       paired = T)
```

##


```
## Paired t-test
##
## data: deeks2_wide %>% pull(cd4_HighVL) and deeks2_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_deeks2 <- 0.0033
# END SOLUTION
```

```
## Test ##
test_that("p4a", {
  expect_true(pvalue_deeks2 > 0 & pvalue_deeks2 < 1)
  print("Checking: range of pvalue_deeks2")
})
```

```
## [1] "Checking: range of pvalue_deeks2"
## Test passed
```

```
## Test ##
test_that("p4b", {
  expect_true(all.equal(pvalue_deeks2, 0.0033, .0001))
  print("Checking: value of pvalue_deeks2")
})
```

```
## [1] "Checking: value of pvalue_deeks2"
## Test passed
```

The paired t-test has a t-statistic of -3.04 for a p-value of 0.003. Assuming the null hypothesis is true (no difference between the mean individual difference at low vs. high viral load), there is only a 0.3% chance of seeing the estimated mean difference we saw in these samples (or a result more extreme). Given that this chance is so low, there appears to be 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 -48.28 and it has a 95% CI of -79.97 to -16.60, 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.

If you are doing this lab before your lab section, please answer the questions using the sample Googlesheet

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

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

6. 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/1Ao2Y9sSwGlgHDct2I4/edit?usp=sharing>

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

Lab 104B: <https://docs.google.com/spreadsheets/d/1ZxsNxSLv514xfHyK3Nlb0dyy8Q4vAjGnH6zZU/edit?usp=sharing>

Lab 105B: <https://docs.google.com/spreadsheets/d/1qAeUPN6PsvVHPgRWrrCEWSRm1blUbzPDAqaVn0B0gNI/edit?usp=sharing>

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

Lab 107B: https://docs.google.com/spreadsheets/d/1z8onu78ZNzv__RlwyYPsnrch8lidt1ABvQDnCg0I9jeQ/edit?usp=sharing

Lab 108B: <https://docs.google.com/spreadsheets/d/1L2e1X7BQBvK5QgjFieOncv0/edit?usp=sharing>

Lab 109B: <https://docs.google.com/spreadsheets/d/1dQes48BgRpt9FjLOeLHmLv0jaiOmBV-fsk/edit?usp=sharing>

Lab 110B: <https://docs.google.com/spreadsheets/d/1pjBrYQG6ObIRmcaRPdEpogE/edit?usp=sharing>

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

Load these packages:

```
library(google Sheets)
```

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

```
library(dplyr)
```

Hit “enter” after your lab. For example, if I am in Lab101B, I would hit “enter” after Lab101B: so the variable `my_key` is on its own line and takes the value of the long string of letters and numbers.

```
#Lab101B: my_key <- "1IxybE5KAHHwLKNni5jFErT8e7JXp1nhYB64kHL6Sheg"
#Lab102B: my_key <- "1Ao2Y9sSwGlgHDct2I4z4c6q6TPETTYTsfp7Uy6o5L0"
#Lab103B: my_key <- "1OmUtQZ79Dx68Gfl8kJgAZrfPN6mAnveDTAIHBUY6tWw"
#Lab104B: my_key <- "1ZxsNwSLv514xfHyK3NYtBF5_lb0dy8Q4vAjGnH6zZU"
#Lab105B: my_key <- "1qAeUPN6PsvVHPgRW-rCEWSRm1blUbzPDAqaVnOB0gNI"
#Lab106B: my_key <- "1rgY7CEtvRUSvVD6mVFRi003u4UbaOfae2mJammCfdEg"
#Lab107B: my_key <- "1z8onu78ZNzv_RlwYYPsnrch8lidtLABvQDnCGOI9jeQ"
#Lab108B: my_key <- "1L2e1X7BQBvK5QgjFieZeiUX0mq6D9ly8eDpVW-0ncv0"
#Lab109B: my_key <- "1dQes48BgRpt9FjLOeLHh2OedqpR_mLv0jaOmBV-fsk"
#Lab110B: my_key <- "1pjBrYQG6ObIRmcaRPnualvpu8isf02N7uoa6-dEpogE"

#sample:
my_key <- "1v9Mvm2hAOB3orINrcbVldJaH1UHcsotyynn91Ppyi70Y"
```

Run these lines of code.

STOP: Remove `eval = F` before continuing

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

```
## Warning: 'arrange()' was deprecated in dplyr 0.7.0.
## Please use 'arrange()' instead.
## See vignette('programming') for more help
```

```
## Worksheets feed constructed with public visibility
```

```
## Warning: 'as_data_frame()' was deprecated in tibble 2.0.0.
## Please use 'as_tibble()' instead.
## The signature and semantics have changed, see '?as_tibble'.
```

```
## Accessing worksheet titled 'Sheet1'.

## Warning: 'select_()' was deprecated in dplyr 0.7.0.
## Please use 'select()' instead.

## Warning: 'filter_()' was deprecated in dplyr 0.7.0.
## Please use 'filter()' instead.
## See vignette('programming') for more help

## Parsed with 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()
## )
```

```
# This creates a dataframe called our_sheet
```

7. 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.

BEGIN QUESTION

name: p7

manual: true

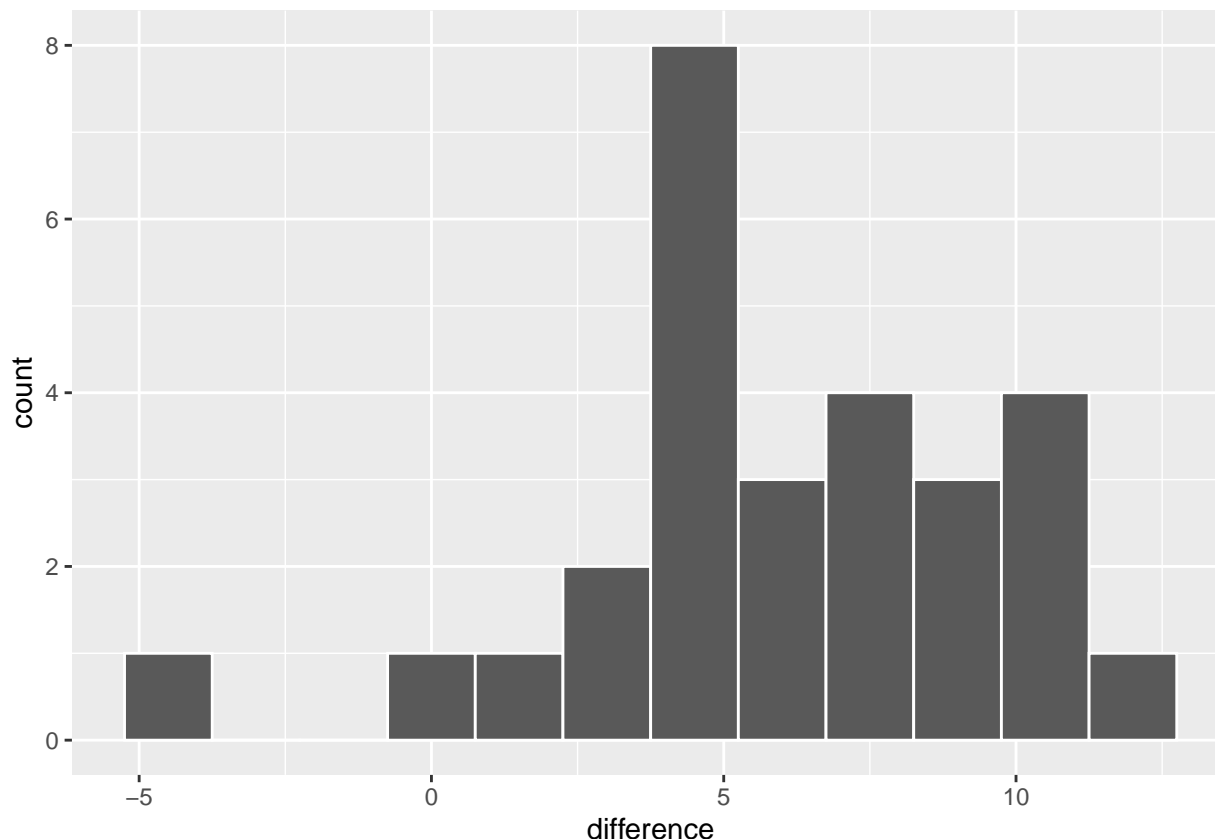
```
. = " # BEGIN PROMPT
## PUT YOUR CODE HERE: Write your code here to investigate the other assumption.
## Hint: You need to first compute a new variable using dplyr before you make your plot :)

# hands <- 'YOUR CODE HERE'
" # END PROMPT

# BEGIN SOLUTION NO PROMPT
hands <- our_sheet %>% mutate(difference = Dominant_num_dots_hit - Non_dominant_num_dots_hit)

ggplot(data = hands, aes(x = difference)) + geom_histogram(binwidth = 1.5, col = "white") # may need to

## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



END *SOLUTION*

- Paired differences in the population are Normally distributed. Can investigate this by looking at the distribution of the paired differences using a histogram or QQ plot. With so little data it is hard to assess Normality, but we can say that there are no really distant outliers based on this sample and using a binwidth of 2 these data look “Normal-ish”.
 - Sample is a SRS from the population. Who is the population here? We didn’t specify! If the population is the ~300 PH142 students can you consider your lab section as a SRS? Why or why not? What about if the population is all students at Berkeley? Even if not a SRS, can we treat it like one given the outcome we’re interested in?
8. 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 `data` to the name of your saved dataset.

STOP: Remove `eval = F` before continuing

BEGIN QUESTION

```
name: p8
manual: true
```

```
# 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 = hands, 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, you can see that most students had a higher number of dots hit with their dominant vs. non-dominant hand, and all but one showed some difference between the two. Since there does seem to be a difference in the number of dots hit, so we would expect that the tests would reject a null hypothesis of no difference.

9. 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.

BEGIN QUESTION

```
name: p9
manual: false
points: 1
```

```
. = " # BEGIN PROMPT
## PUT YOUR T-TEST CODE HERE

pvalue_paired <- 'REPLACE WITH YOUR PVALUE ROUNDED TO 2 DECIMAL PLACES'
" # END PROMPT

# BEGIN SOLUTION NO PROMPT
t.test(hands %>% pull(Dominant_num_dots_hit),
       hands %>% pull(Non_dominant_num_dots_hit),
       paired = T, alternative = "two.sided")

##
## Paired t-test
##
## data: hands %>% pull(Dominant_num_dots_hit) and hands %>% pull(Non_dominant_num_dots_hit)
## t = 8.9753, df = 27, p-value = 1.369e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.573246 7.283897
## sample estimates:
## mean of the differences
##          5.928571

pvalue_paired <- 0.00
# END SOLUTION
```

```
## Test ##
test_that("p9a", {
  expect_true(pvalue_paired > -0.01 & pvalue_paired < 1)
  print("Checking: range of pvalue_paired")
})
```

```
## [1] "Checking: range of pvalue_paired"
## Test passed
```

```
## Test ##
test_that("p9b", {
  expect_true(all.equal(pvalue_paired, 0.00, .01))
  print("Checking: value of pvalue_paired")
})
```

```
## [1] "Checking: value of pvalue_paired"
## Test passed
```

The p-value for the test is close to 0 (basically 0). 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 that we saw (or one of greater magnitude) between the sample average of the individual differences. This probability is so small that we reject the null hypothesis in favour of the alternative hypothesis that there is a difference between the hands in the number of dots hit.

The 95% CI for the difference is 3.49 to 6.02. If our model assumptions are correct and there is only random error affecting the estimate, this method for calculating confidence intervals will contain the true underlying value of difference between dominant and non-dominant hands 95% of the time (19 times out of 20).

10. 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?

BEGIN QUESTION

```
name: p10
manual: false
points: 1
```

```
. = " # BEGIN PROMPT
# YOUR T-TEST CODE HERE

# Then, uncomment one of these choices:
# p10 <- 'smaller'
# p10 <- 'larger'
" # END PROMPT

# BEGIN SOLUTION NO PROMPT
t.test(hands %>% pull(Dominant_num_dots_hit),
       hands %>% pull(Non_dominant_num_dots_hit),
       paired = F, alternative = "two.sided")
```

```
##
## Welch Two Sample t-test
```

```
##
## data:  hands %>% pull(Dominant_num_dots_hit) and hands %>% pull(Non_dominant_num_dots_hit)
## t = 5.9868, df = 53.088, p-value = 1.89e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.942422 7.914721
## sample estimates:
## mean of x mean of y
## 28.25000 22.32143
```

```
p10 <- "larger"
# END SOLUTION
```

```
## Test ##
test_that("p10", {
  expect_true(all.equal(p10, "larger"))
  print("Checking: selection")
})
```

```
## [1] "Checking: selection"
## Test passed
```

Solution A (based on specific data): Using these data, the p-value is larger. Here the p-value was still a very small number, and still provides evidence against the null hypothesis of no difference between the groups. This is because the mean number of dots hit for the dominant hand (28.4) was so much larger than for the non-dominant hand (22.6) that the test still detected the difference, even though we forgot to tell the test that these data were paired.

Solution B (generic): The p-value for the independent test is larger. This means there is less evidence (or maybe even minimal/“no”) evidence against the null hypothesis of no difference. It is larger/there is less evidence because the independent test does not account for the matching in the data. So it included all the “noise” due to differences between individuals.

11. 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?

BEGIN QUESTION

```
name: p11
manual: true
```

Solution B: It could be that there was a “learning effect” and that when you played the game with the first hand, you partially struggled because it was your first time playing the game. If that was the case, then you would expect to do a bit better using your second hand because you are more familiar with the game. To investigate this, we could have made a histogram of the individual differences for those who used their dominant hand first, and compared it to a separate histogram for those who used their non-dominant hand first. If these plots look very different this might suggest that there was a learning effect.

Solution A: It could be that playing the game had a “learning” effect, meaning that you learnt how to play it with the first hand, which means that you become better at playing it with the second hand. If this was the case, then those that use their dominant hand second are expected to perform better with that hand than those who used their dominant hand first. To investigate this. We could make a boxplot for the number of dots hit with the first hand, and a second box plot for the second hand, and do this *separately* according to which hand was first.

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-sp21/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.