Assignment 7

Felicia Liu

10/30/2020

- Solutions released date: Tuesday, November 3.

Helpful hints:

- Every function you need to use was taught during lecture! So you may need to revisit the lecture code to help you along by opening the relevant files on Datahub.

In two wards for elderly patients in a local hospital the following levels of hemoglobin (grams per liter) were found for a simple random sample of patients from each ward.:

Ward A:

```
ward_a <- c(12.2, 11.1, 14.0, 11.3, 10.8, 12.5, 12.2, 11.9, 13.6, 12.7, 13.4, 13.7)
```

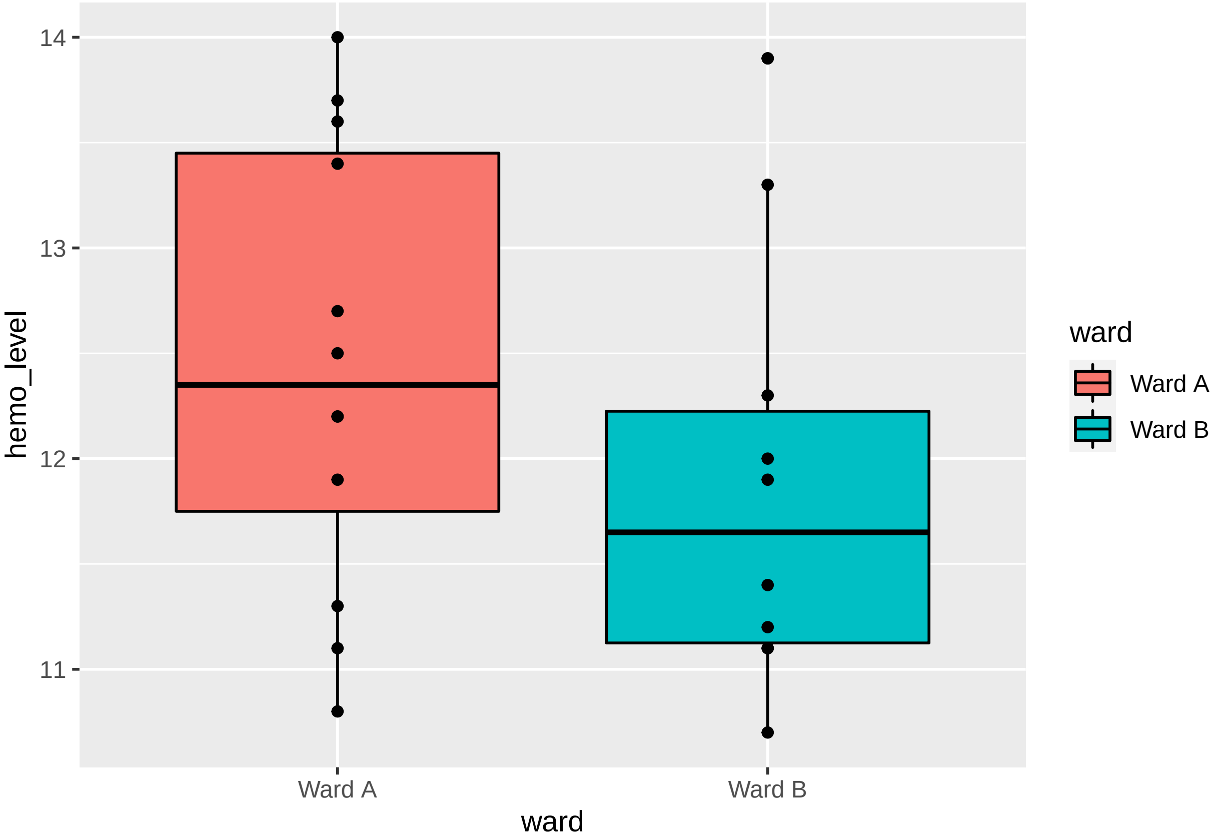
Ward B:

```
ward_b <- c(11.9, 10.7, 12.3, 13.9, 11.1, 11.2, 13.3, 11.4, 12.0, 11.1)
```

1. [1 point] In one ggplot, create two box plots to compare the hemoglobin values for Ward A and Ward B. Also plot the raw data as points, overlaid on top of the box plots.

```
hemoglobin <- data.frame(hemo_level = c(ward_a, ward_b),
                         ward = c(rep("Ward A", 12), rep("Ward B", 10)))
```

```
p1 <- ggplot(hemoglobin, aes(x = ward, y = hemo_level)) +
  geom_boxplot(aes(fill = ward), col = "black") + geom_point()
p1
```



```
check_problem1()
```

```
## [1] "Checkpoint 1 Passed: A ggplot has been defined"
## [1] "Checkpoint 2 Passed: Correct! You used hemoglobin as the data."
## [1] "Checkpoint 3 Passed: Correct! You used hemoglobin as the data!"
## [1] "Checkpoint 4 Passed: Correct! You used hemoglobin as the data!"
## [1] "Checkpoint 5 Passed: Correct! You defined a boxplot."
## [1] "Checkpoint 6 Passed: Correct! You plotted the raw data points with geom_point overlaid on top o
##
## Problem 1
## Checkpoints Passed: 6
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

2. [1 points] Comment on the similarities/differences portrayed by the plots, keeping in mind that the sample size is relatively small for these two wards.

The minimum values of the boxplots are relatively similar. However, the first quartile, median, third quartile value, and maximum are higher in Ward A than in Ward B. Both samples appear to be roughly symmetric and there is some overlap in the middle 50% of data.

3. [2 points] What two assumptions do you need to make to use any of the t-procedures? Because each ward has a rather small sample size ($n < 12$ for both), what two characteristics of the data would you need to check for to ensure that the t-procedures can be applied?

The sample is randomly chosesn (i.e., an SRS) and independent, and the distribution of the populations are Normal. There should be no outliers and data should be similar shapes.

4. [3 points] Using only dplyr and *t functions, create a 95% confidence interval for the mean difference between Ward A and Ward B. You can do this by using dplyr to calculate the inputs required to calculate the 95% CI, and then plugging these values in on a separate line of code (or using your calculator). Use a degrees of freedom of 19.515 (You don't need to calculate the degrees of freedom, you can use this value directly). Show your work and interpret the mean difference and its 95% CI. Round your solution to 3 decimal places.

```
hemoglobin %>% group_by(ward) %>%
              summarize(mean_hemo = mean(hemo_level),
              sample_var = var(hemo_level), n = length(hemo_level))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 2 x 4
##   ward   mean_hemo sample_var     n
##   <chr>      <dbl>      <dbl> <int>
## 1 Ward A      12.4       1.14    12
## 2 Ward B      11.9       1.07    10
```

```
deg_free <- ( (1.140909/12) + (1.065444/10) )^2/( (1/11)*(1.140909/12)^2 + (1/9)*(1.065444/10)^2)
mean_diff <- 12.45 - 11.89
se_diff <- sqrt(1.140909/12 + 1.065444/10)
t_star <- qt(p = 0.025, df = deg_free)


# THEN, ASSIGN YOUR FINAL ANSWERS BELOW:
CI_lowerbound <- round(mean_diff - t_star * (se_diff), 3)
CI_upperbound <- round(mean_diff + t_star * (se_diff), 3)
#c(CI_lowerbound, CI_upperbound) %>% round(3)

check_problem4()
```

```
## [1] "Checkpoint 1 Error: Did you use mean_diff - t_star * (se_diff)?"
## [1] "Checkpoint 2 Error: Did you use mean_diff + t_star * (se_diff)?"
##
## Problem 4
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

5. [1 points] Interpret the mean difference and its 95% CI you just calculated.

There is a 95% chance that the true mean difference value lies between 0.44 and 1.56.

Perform a two-sided t-test for the difference between the two samples, where the null hypothesis is that the underlying means are the same. Start by writing down the null and alternate hypotheses, then calculate the test statistic (showing your work) and p-value. Continue to assume that the degrees of freedom is 19.515. Verify the p-value by running the t-test using R's built in function. Show the output from that test. Hint: to perform the t-test using R's built in function, you need to pass the function an x and y argument, where x includes that values for Ward A and Y includes the values for Ward B. `dplyr`'s `filter()` and `pull()` functions will be your friends.

6. [1 points] Calculate the t-test statistics

```
t.test(hemoglobin %>% filter(ward == "Ward A") %>% pull(hemo_level),
                 hemoglobin %>% filter(ward == "Ward B") %>% pull(hemo_level),
                 alternative = "two.sided", mu = 0)
```

```
##
##  Welch Two Sample t-test
##
## data:  hemoglobin %>% filter(ward == "Ward A") %>% pull(hemo_level) and hemoglobin %>% filter(ward =
## t = 1.2472, df = 19.515, p-value = 0.2271
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3781372  1.4981372
## sample estimates:
## mean of x mean of y
##     12.45     11.89
```

```
t_statistics <- 1.2472
t_statistics
```

```
## [1] 1.2472
```

```
check_problem6()
```

```
## [1] "Checkpoint 1 Passed: Correct t statistic!"
##
## Problem 6
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

7. [1 points] We need to compare this t-statistic to a t distribution with 19.515 degrees of freedom:

```
p_value <- pt(1.247157, df = 19.515, lower.tail = F) * 2
p_value
```

```
## [1] 0.2271006
```

```
check_problem7()
```

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

8.[2 points] Interpret the p value you got in the context of the this question. Are there evidence against null hypothesis?

The p-value of 0.2271 is relatively large and means that we fail to reject the null hypothesis, and there is not enough evidence against the null hypothesis. There is a 22.7% chance of seeing a difference of the size we saw or larger under the hypothesis of no difference.

Check this against the `t.test` output:

```
t.test(x = hemoglobin %>% filter(ward == "Ward A") %>% pull(hemo_level),
       y = hemoglobin %>% filter(ward == "Ward B") %>% pull(hemo_level),
       alternative = "two.sided")
```

```
## 
##  Welch Two Sample t-test
## 
## data:  hemoglobin %>% filter(ward == "Ward A") %>% pull(hemo_level) and hemoglobin %>% filter(ward =
## t = 1.2472, df = 19.515, p-value = 0.2271
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3781372  1.4981372
## sample estimates:
## mean of x mean of y 
##     12.45     11.89
```

The time to perform open heart surgery is normally distributed. Sixteen patients (chosen as a simple random sample from a hospital) underwent open heart surgery that took the following lengths of time (in minutes).

```r
op_time <- c(247.8648, 258.4343, 315.6787, 268.0563, 269.9372, 320.6821,
             280.5493, 225.3180, 243.8207, 251.5388, 304.9706, 277.3140,
             278.6247, 269.3418, 248.0131, 322.9812)
surg_data <- data.frame(op_time)
```

9. [1 point] You wish to know if the mean operating time of open heart surgeries at this hospital exceeds four hours. Set up appropriate hypotheses for investigating this issue.

The null hypothesis is that the open heart surgery takes 4 hours to complete. The alternative hypothesis is that the open heart surgery takes more than 4 hours to complete.

10. [1 points] Test the hypotheses you formulated in part (a). Report the p-value. (Do not use the `t.test` function for this question)

```
surg_data %>% summarize (sample_mean = mean(op_time),
              sample_sd = sd(op_time),
              sample_size = length(op_time),
              sample_se = sample_sd/sqrt(sample_size),
              t_test = (sample_mean - 4)/sample_se,
              p_value = 1 - pt(t_test, df = sample_size - 1))
```

```
##   sample_mean sample_sd sample_size sample_se   t_test      p_value
## 1    273.9454  29.22249          16  7.305622 36.95036 2.220446e-16
```

```
p_value_10 <- 2.220446e-16
p_value_10
```

```
## [1] 2.220446e-16
```

```
check_problem10()
```

```
## [1] "Checkpoint 1 Passed: Correct!"
##
## Problem 10
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

11. [1 points] What are your conclusions in the context of the question?

With a p-value of 2.220446e-16, which is extremely small, we are able to reject the null hypothesis that the open heart surgeries at this hospital take 4 hours to complete. We have evidence against the null hypothesis.

12. [3 points] Construct a 95% CI on the mean operating time (in hours).

```
#t.test(x = surg_data %>% pull(op_time), alternative = "greater", mu = 4)
qt(p = 0.975, df = 15)
```

```
## [1] 2.13145
```

```
# THEN, ASSIGN YOUR ANSWERS BELOW:
CI_lowerbound_12 <- 4.31
CI_upperbound_12 <- 4.73

check_problem12()
```

```
## [1] "Checkpoint 1 Passed: Correct!"
## [1] "Checkpoint 2 Passed: Correct"
##
## Problem 12
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

13. [1 point] Suppose you were testing the hypotheses $H_0 : \mu_d = 0$ and $H_a : \mu_d \neq 0$ in a paired design and obtain a p-value of 0.21. Which one of the following could be a possible 95% confidence interval for μ_d?

```r
# Uncomment one of the following choices:
# p13 <- "-2.30 to -0.70"
p13 <- "-1.20 to 0.90"
# p13 <- "1.50 to 3.80"
# p13 <- "4.50 to 6.90"

check_problem13()
```

```
## [1] "Checkpoint 1 Passed: Correct!"
## 
## Problem 13
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

14. [1 point] Suppose you were testing the hypotheses $H_0 : \mu_d = 0$ and $H_a : \mu_d \neq 0$ in a paired design and obtain a p-value of 0.02. Also suppose you computed confidence intervals for μ_d. Based on the p-value which one of the following is true?

```
# Uncomment one of the following choices:
# p14 <- "Both a 95% CI and a 99% CI will contain 0."
# p14 <- "A 95% CI will contain 0, but a 99% CI will not."
p14 <- "A 95% CI will not contain 0, but a 99% CI will."
# p14 <- "Neither a 95% CI nor a 99% CI interval will contain 0."

check_problem14()
```

```
## [1] "Checkpoint 1 Passed: Correct!"
##
## Problem 14
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

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  NOT YET GRADED               1 free-response
## Problem 3  NOT YET GRADED               2 free-response
## Problem 4          FAILED               3    autograded
## Problem 5  NOT YET GRADED               1 free-response
## Problem 6          PASSED               1    autograded
## Problem 7          PASSED               1    autograded
## Problem 8  NOT YET GRADED               2 free-response
## Problem 9  NOT YET GRADED               1 free-response
## Problem 10         PASSED               1    autograded
## Problem 11 NOT YET GRADED               1 free-response
## Problem 12         PASSED               3    autograded
## Problem 13         PASSED               1    autograded
## Problem 14         PASSED               1    autograded
```