Homework 10

Felicia Liu

11/06/2020

- Solutions released date: Tuesday, October 27
- Remember: autograder is meant as sanity check ONLY. It will not tell you if you have the correct answer. It will tell you if you are in the ball park of the answer so *CHECK YOUR WORK*.
- 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.

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. Alternatively, you may wish to view the code in the condensed PDFs posted on the course website. Good luck!

- Knit your file early and often to minimize knitting errors! If you copy and paste code for the slides, you are bound to get an error that is hard to diagnose. Typing out the code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily. This will save you and the GSIs from frustration! **You must knit correctly before submitting.**

- If your code runs off the page of the knitted PDF then you will LOSE POINTS! To avoid this, have a look at your knitted PDF and ensure all the code fits in the file (you can easily view it on Gradescope via the provided link after submitting). If it doesn't look right, go back to your .Rmd file and add spaces (new lines) using the return or enter key so that the code runs onto the next line.

You would like to conduct a survey of highschool students to determine the proportion who are current e-cigarettes users. Before you conduct your survey, you need to determine how large of a sample size. Suppose that you would like the width of the 95% confidence interval to be 2.5 percentage points.

1. [1 point] Determine the most conservative sample size you would require and assign it to object p1. Recall that to do this, you need to use a p^* of 0.5.

```r
p1 <- ceiling((1.96/0.025)**2 * 0.5 * (1 - 0.5))
# remember to remove " " if you want to store a number


check_problem1()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric"
## [1] "Checkpoint 2 Passed: Correct, the answer is 1537."
##
## Problem 1
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

2. [1 point] You've seen a recent publication from the Annals of Internal Medicine that estimated that 9.2% of individuals aged 18 to 24 years old are current e-cigarette users. What is the sample size estimate assuming that $p^* = 0.092$.

```
p2 <- ceiling((1.96/0.025)**2 * 0.092 * (1 - 0.092))
# remember to remove " " if you want to store a number


check_problem2()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric"
## [1] "Checkpoint 2 Passed: Correct, the answer is 514."
##
## Problem 2
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

3. [1 point] The recent publication referenced in the previous question only looked at adults (aged 18+), but observed that the rate of current use was inversely related to age among the population they surveyed. Because of this finding would you suppose that the sample size estimated in part (b) is too low or too high?

```
# Uncomment one of the following options:
p3 <- "Too low"
# p3 <- "Too high"


check_problem3()
```

```
## [1] "Checkpoint 1 Passed: Correct"
## [1] "Checkpoint 2 Passed: Correct, the sample size estimated in part (b) is too low."
##
## Problem 3
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Exclusive breastfeeding during the first six months of life is recommended for optimal infant growth and development. Suppose that you conducted a survey of randomly chosen women from California and found that 775 out of 5615 new mothers exclusively breast fed their infants.

Perform all four of the methods discussed in lecture and during lab to create a 95% confidence interval for the proportion of infants exclusively breast fed.

```r
library(tidyverse)
library(tibble)
```

Store your answer to p4-p7 using the following format:

```r
#pX <- c(lowerbound, upperbound)

# For example, if lowerbound = 10, upperbound = 20:
pX <- c(10, 20)
```

4. [1 point] Use the large sample method of constructing a 95% CI.

```r
p.hat <-  775/5615
se <- sqrt(p.hat * (1 - p.hat)/5615)
p.hat - 1.96*se
```

```
## [1] 0.1290011
```

```r
p.hat + 1.96*se
```

```
## [1] 0.1470452
```

```r
# Replace "lowerbound" and "upperbound" with your answer
# If your answer is a number, make sure it doesn't have quotes around it
p4 <- c(0.1290, 0.1470)


check_problem4()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric."
## [1] "Checkpoint 2 Passed: Upperbound and lowerbound have been stored."
## [1] "Checkpoint 3 Passed: Correct, the lowerbound is 0.1290011 and upperbound 0.1470452."
##
## Problem 4
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

5. [1 point] Use the Clopper Pearson (Exact) method of constructing a 95% CI.

```r
binom.test(x = 775, n = 5615, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  775 and 5615
## number of successes = 775, number of trials = 5615, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1291020 0.1473222
## sample estimates:
## probability of success
##              0.1380232
```

```r
# Replace "lowerbound" and "upperbound" with your answer
# If your answer is a number, make sure it doesn't have quotes around it
p5 <- c(0.1291, 0.1473)

check_problem5()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric."
## [1] "Checkpoint 2 Passed: Upperbound and lowerbound have been stored."
## [1] "Checkpoint 3 Passed: Correct, the lowerbound is 0.1291020 and upperbound 0.1473222."
##
## Problem 5
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

6. [1 point] Use the Wilson Score method of constructing a 95% CI with a continuity correction.

```r
prop.test(x = 775, n = 5615, conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  775 out of 5615, null probability 0.5
## X-squared = 2941.4, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1291619 0.1473842
## sample estimates:
##         p
## 0.1380232
```

```r
# Replace "lowerbound" and "upperbound" with your answer
# If your answer is a number, make sure it doesn't have quotes around it
p6 <- c(0.1292,  0.1474)

check_problem6()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric."
## [1] "Checkpoint 2 Passed: Upperbound and lowerbound have been stored."
## [1] "Checkpoint 3 Passed: Correct, the lowerbound is 0.1291619 and upperbound 0.1473842."
##
## Problem 6
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

7. [1 point] Use the Plus Four method of constructing a 95% CI.

```r
p.tilde <- (775 + 2)/(5615 + 4)
se <- sqrt(p.tilde * (1 - p.tilde)/5619)
p.tilde - 1.96 * se
```

```
## [1] 0.1292549
```

```r
p.tilde + 1.96 * se
```

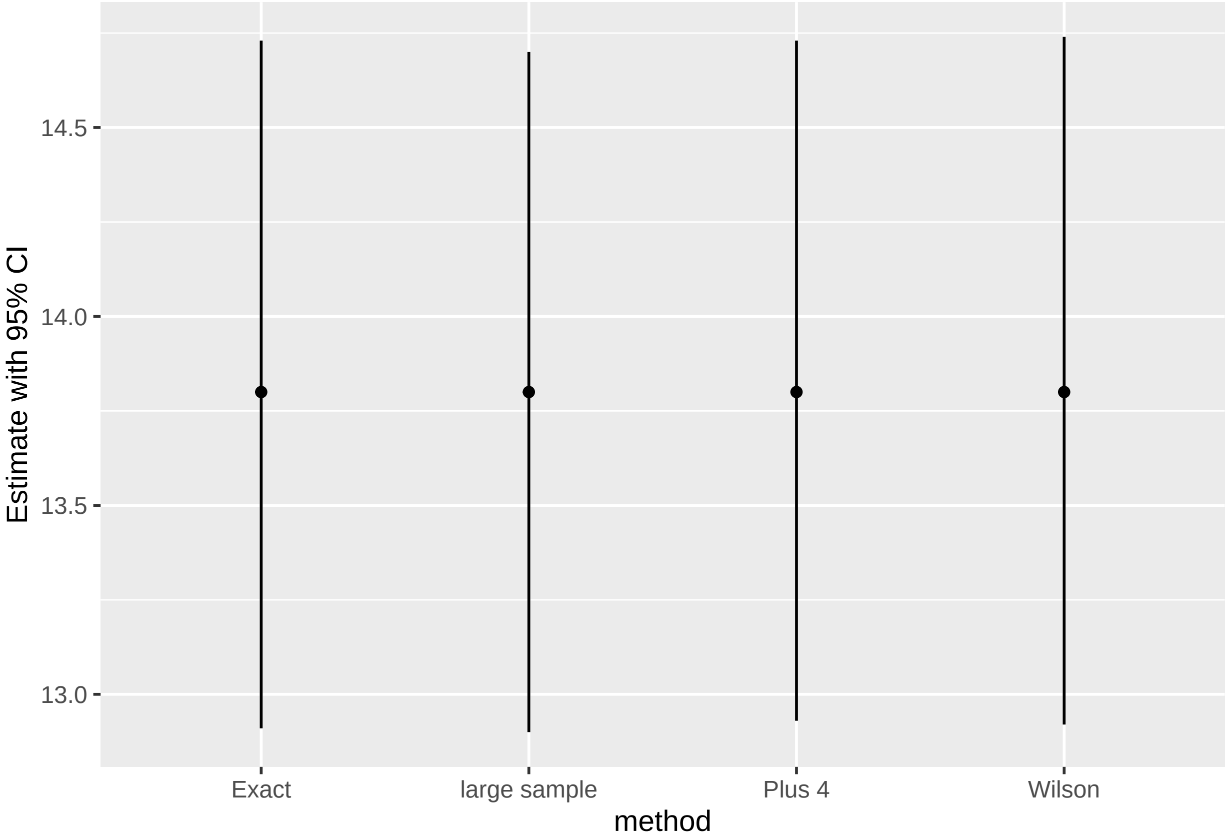
```
## [1] 0.1473067
```

```r
# Replace "lowerbound" and "upperbound" with your answer
# If your answer is a number, make sure it doesn't have quotes around it
p7 <- c(0.1293, 0.1473)

check_problem7()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric."
## [1] "Checkpoint 2 Passed: Upperbound and lowerbound have been stored."
## [1] "Checkpoint 3 Passed: Correct, the lowerbound is 0.1292517 and upperbound 0.1473099."
##
## Problem 7
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

8. [2 points] Create a plot comparing the confidence intervals. If you are stuck, refer back to the example code presented in Lab 8.

```r
library(ggplot2)
library(tibble)
breastfeed_CIs <- tibble(method   = c("large sample", "Exact", "Wilson", "Plus 4"),
                   lower_CI = c(12.90           , 12.91    ,  12.92    ,  12.93),
                   upper_CI = c(14.70           , 14.73    ,  14.74    ,  14.73),
                   estimate = c(13.80           , 13.80    ,  13.80    ,  13.80)
                   )
p8 <- ggplot(data = breastfeed_CIs, aes(x = method, y = estimate)) +
  geom_point() +
  geom_segment(aes(x = method, xend = method, y = lower_CI, yend = upper_CI)) +
  labs(y = "Estimate with 95% CI")

p8
```



```r
check_problem8()
```

```
## [1] "Checkpoint 1 Passed: A ggplot has been defined."
## [1] "Checkpoint 2 Passed: breastfeed_CIs has been used as the data."
## [1] "Checkpoint 3 Passed: The y has been specified correctly."
## [1] "Checkpoint 4 Passed: The x has been specified correctly."
## [1] "Checkpoint 5 Passed: The point estimates have been added."
## [1] "Checkpoint 6 Passed: The CI segments have been added."
##
```

```
## Problem 8
## Checkpoints Passed: 6
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

9. [1 point] Do the methods produce confidence intervals that are basically the same or very different? Why?

The sample size is large enough that the Central Limit Theorem holds, so the confidence interval estimates are basically the same. When the sample size is large enough, the different confidence intervals should be approximately the same.

10. [1 point] Suppose that in 2010, the rate of exclusive breastfeeding in California was known to be 18.6%. Based on the 95% CIs calculated in questions 4-7, is there evidence against the null hypothesis that the underlying rate is equal to 18.6% in favor of the alternative that the rate is different from 18.6%?

18.6% is higher than any of the confidence intervals that we calculated. This means that the p-value is $<$ 0.05 and therefore there is evidence against the null hypothesis and in favor of the alternative hypothesis that the rate differs from 18.6%

To confirm your answer to Problem 9, perform a two-sided hypothesis test and interpret the p-value.

11. [1 point] State the null and alternative hypotheses:

The null hypothesis is mu is equal to 18.6% and the alternative hypothesis is mu is not equal to 18.6%.

12. [1 point] Calculate the test statistic:

```r
p12 <-((p.hat - 0.186)/sqrt(0.186 * (1 - 0.186)/5615))

check_problem12()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric"
## [1] "Checkpoint 2 Passed: Correct, the answer is 9.239275."
##
## Problem 12
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

13. [1 point] Calculate the p-value:

```
p13 <- pnorm(p12, lower.tail = T) * 2


check_problem13()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric"
## [1] "Checkpoint 2 Passed: Correct, the answer is 2.48172e-20."
##
## Problem 13
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

14. [1 point] Interpret the p-value:

A p-value of 2.48172e-20 means that we should reject the null hypothesis of mu equals 18.6% in favor of the alternative, where mu is not equal to 18.6%. The chance of seeing a proportion of 13.8% (or one even more different in magnitude) from the null value of 18.6% is $<$ 0.0001%.

The quadrivalent HPV vaccine was introduced to Canada in 2007, and was given to girls in Ontario, Canada who were enrolled in grade 8 (13-14 year olds). Before 2007, no girls recieved the vaccine, while in the 4 years after it was introduced nearly 40% of girls recieved the vaccine each year. One concern that some people had was that the vaccine itself would increase promiscuity if the girls felt a false sense of protection, which could thereby increase the prevalence of other sexually transmitted infections (STIs) among vaccinated girls. This paper examines this question using an advanced method called the "regression discontinuity" design which harnesses the abrupt change in vaccination status across cohorts of girls to estimate the causal effect of vaccination against HPV on the occurrence of other STIs.

Read only the abstract of the paper, and don't worry about the details because these are advanced methods. Note that the term "RD" is the difference in risk of STIs between girls exposed and unexposed to HPV vaccination. We can therefore think of this risk difference as the difference between two proportions.

15. [1 point] Interpret this result from the abstract: We identified 15 441 (5.9%) cases of pregnancy and sexually transmitted infection and found no evidence that vaccination increased the risk of this composite outcome: RD per 1000 girls -0.61 (95% confidence interval [CI] -10.71 to 9.49).

In particular, what does -0.61 estimate?

-0.61 is the estimated difference in the proportions of girls with an STI comparing girls who were vaccinated and girls who were not vaccinated.

16. [1 point] The 95% confidence interval includes 0. What can you conclude about the p-value for a two-sided test of the difference between vaccinated and unvaccinated girls and their risk of sexually transmitted diseases?

Given that the null value for the H0 that there is no difference is included in the 95% CI, we know that the corresponding two-sided test of the difference between the underlying proportions would be greater than 5%.

An allergy to peanuts is increasingly common in Western countries. A randomized controlled trial enrolled infants with a diagnosed peanut sensitivity. Infants were randomized to either avoid peanuts or to consume them regularly until they reached age 5. At the end of the study, 18 out of the 51 randomized to avoid peanuts were tested to be allergic to peanuts. Only 5 out of the 47 randomized to consuming them regularly were tested to be allergic to peanuts.

17. [1 point] Estimate the difference between the two proportions.

```
p17 <- 18/51 - 5/47


check_problem17()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric"
## [1] "Checkpoint 2 Passed: Correct, the answer is 0.2465582."
##
## Problem 17
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

18. [1 point] Use the plus four method to find a 99% confidence interval for the difference between the two groups. Store the upper and lower bounds into an object called p18.

```
p.tilde1 = (18+1)/(51+2)
p.tilde2 = (5+1)/(47+2)
se <- sqrt((p.tilde1 * (1 - p.tilde1)/(51 + 2)) + (p.tilde2 * (1 - p.tilde2)/47 + 2))
(p.tilde1 - p.tilde2) - 2.576 * se
```

```
## [1] -3.413002
```

```
(p.tilde1 - p.tilde2) + 2.576 * se
```

```
## [1] 3.885085
```

```
# Replace "lowerbound" and "upperbound" with your answer
# If your answer is a number, make sure it doesn't have quotes around it
p18 <- c((p.tilde1 - p.tilde2) - 2.576 * se, (p.tilde1 - p.tilde2) + 2.576 * se)


check_problem18()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric."
## [1] "Checkpoint 2 Passed: Upperbound and lowerbound have been stored."
## [1] "Checkpoint 3 Error: Wrong answer. Incorrect value of the interval."
##
## Problem 18
## Checkpoints Passed: 2
## Checkpoints Errored: 1
## 66.67% passed
## --------
## Test: FAILED
```

19. [1 point] Why would it have been inappropriate to use the large sample method to create a 99% CI?

The number of successes out of the kids who ate peanuts was 5, and $5 < 10$, so the large sample method is not appropriate to create the interval.

Perform a two-sided hypothesis test for the difference between the groups. Start by stating the null and alternative hypotheses, then calculate the test statistic, the p-value, and conclude with your interpretation of the p-value.

20. [1 point] State the null and alternative hypotheses:

Null hypothesis: p1 = p2

Alternative hypothesis: p1 is not equal to p2

21. [1 point] Calculate the test statistic:

```
p.hat <- (18+5)/(51+47)
p21 <- (18/51 - 5/47)/sqrt(p.hat * (1 - p.hat) * (1/51 + 1/47))


check_problem21()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric"
## [1] "Checkpoint 2 Passed: Correct, the answer is 2.877213."
##
## Problem 21
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

22. [1 point] Calculate the p-value:

```
p22 <- pnorm(p21, lower.tail = F) * 2


check_problem22()
```

```
## [1] "Checkpoint 1 Passed: Your answer is numeric"
## [1] "Checkpoint 2 Passed: Correct, the answer is 0.004012052."
##
## Problem 22
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

23. [1 point] Interpret the p-value:

A p-value of 0.004012 is very small and means that we can reject the null hypothesis of p1 = p2 in favor of the alternative hypothesis, p1 is not equal to p2.

24. [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?

```
# Uncomment one of the following choices:
# p24 <- "-2.30 to -0.70"
p24 <- "-1.20 to 0.90"
# p24 <- "1.50 to 3.80"
# p24 <- "4.50 to 6.90"


check_problem24()
```

```
## [1] "Checkpoint 1 Passed: One choice has been selected."
## [1] "Checkpoint 2 Passed: Correct"
##
## Problem 24
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

25. [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 of the following are true?

```
# Uncomment one of the following choices:
# p25 <- "Both a 95% CI and a 99% CI will contain 0."
# p25 <- "A 95% CI will contain 0, but a 99% CI will not."
p25 <- "A 95% CI will not contain 0, but a 99% CI will."
# p25 <- "Neither a 95% CI nor a 99% CI interval will contain 0."


check_problem25()
```

```
## [1] "Checkpoint 1 Passed: One choice has been selected."
## [1] "Checkpoint 2 Passed: Correct"
##
## Problem 25
## Checkpoints Passed: 2
## 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.

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

```
##                       Test Points_Possible          Type
## Problem 1           PASSED               1    autograded
## Problem 2           PASSED               1    autograded
## Problem 3           PASSED               1    autograded
## Problem 4           PASSED               1    autograded
## Problem 5           PASSED               1    autograded
## Problem 6           PASSED               1    autograded
## Problem 7           PASSED               1    autograded
## Problem 8           PASSED               2    autograded
## Problem 9  NOT YET GRADED               1 free-response
## Problem 10 NOT YET GRADED               1 free-response
## Problem 11 NOT YET GRADED               1 free-response
## Problem 12          PASSED               1    autograded
## Problem 13          PASSED               1    autograded
## Problem 14 NOT YET GRADED               1 free-response
## Problem 15 NOT YET GRADED               1 free-response
## Problem 16 NOT YET GRADED               1 free-response
## Problem 17          PASSED               1    autograded
## Problem 18          FAILED               1    autograded
## Problem 19 NOT YET GRADED               1 free-response
## Problem 20 NOT YET GRADED               1 free-response
## Problem 21          PASSED               1    autograded
## Problem 22          PASSED               1    autograded
## Problem 23 NOT YET GRADED               1 free-response
## Problem 24          PASSED               1    autograded
## Problem 25          PASSED               1    autograded
```