

# Problem Set 10

Your name and student ID

Today's date

**Run this chunk of code to load the autograder package!**

## Instructions

- Solutions will be released by Sunday, Aug 3rd.
- This semester, homework assignments are for practice only and will not be turned in for marks.

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!
- It is good practice to not allow your code to run off the page. To avoid this, have a look at your knitted PDF and ensure all the code fits in the file. 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 high school 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 you will need. Suppose that you would like the width of the 95% confidence interval to be 5 percentage points.

**1. Determine the most conservative sample size you would need and assign this value to the object p1. Hint: you will need to use a  $p^*$  of 0.5.**

```
p1 <- ceiling((1.96/0.025)^2*0.5*(1-0.5))
p1
```

```
## [1] 1537
```

```
. = ottr::check("tests/p1.R")
```

```
##
```

```
## All tests passed!
```

$$n = (z^*/m)^2 p^* (1 - p^*) \quad n = (1.96/0.025)^2 \times 0.5 \times (1 - 0.5) = 1536.64 = 1537$$

Thus, we would need a sample size of 1537 high school students to obtain a margin of error of 2.5 percentage points if we assume the true prevalence is 50%.

2. Next, let's say 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))  
p2
```

```
## [1] 514
```

```
. = ottr::check("tests/p2.R")
```

```
##
```

```
## All tests passed!
```

$$n = (z^*/m)^2 p^* (1 - p^*) \quad n = (1.96/0.025)^2 \times 0.092 \times (1 - 0.092) = 513.459 = 514$$

Thus, we would need a sample size of 514 high school students to obtain a margin of error of 2.5 percentage points if we assume the true prevalence is 9.2%.

3. The 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 you estimated in question 2 is too low or too high?

```
p3 <- 'Too low'
```

I would suppose that the estimated sample size is too low because the true prevalence among highschool students is likely higher than among 18-24 year olds. If that is the case, then using a higher  $p^*$  in the sample size calculation would increase the sample size required.

```
. = ottr::check("tests/p3.R")
```

```
##
```

```
## All tests 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 breastfed 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 breastfed.

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

#### 4. Construct a 95% confidence interval using the large sample method.

```
num_successes <- 775
sample_size <- 5615

p_hat <- num_successes/sample_size # estimate proportion
se <- sqrt(p_hat*(1-p_hat)/sample_size) # standard error
p4 <- c(p_hat - 1.96*se, p_hat + 1.96*se) # CI
p4
```

```
## [1] 0.1290011 0.1470452
```

```
. = ottr::check("tests/p4.R")
```

```
##
## All tests passed!
```

#### 5. Construct a 95% confidence interval using the Clopper Pearson (Exact) method.

```
exact_out <- binom.test(num_successes, sample_size, p=0.5)
p5 <- c(exact_out$conf.int[1], exact_out$conf.int[2])
p5
```

```
## [1] 0.1291020 0.1473222
```

```
. = ottr::check("tests/p5.R")
```

```
##
## All tests passed!
```

#### 6. Construct a 95% confidence interval using the Wilson Score method with the continuity correction.

```
wilson_out <- prop.test(num_successes, sample_size, p=0.5)
p6 <- c(wilson_out$conf.int[1], wilson_out$conf.int[2])
p6
```

```
## [1] 0.1291619 0.1473842
```

```
. = ottr::check("tests/p6.R")
```

```
##
```

```
## All tests passed!
```

**7. Construct a 95% confidence interval using the Plus Four method.**

```
p_tilde <- (num_successes + 2)/(sample_size + 4)
se <- sqrt(p_tilde*(1 - p_tilde)/sample_size) # standard error
p7 <- c(p_tilde - 1.96*se, p_tilde + 1.96*se) # CI
p7
```

```
## [1] 0.1292517 0.1473099
```

```
. = ottr::check("tests/p7.R")
```

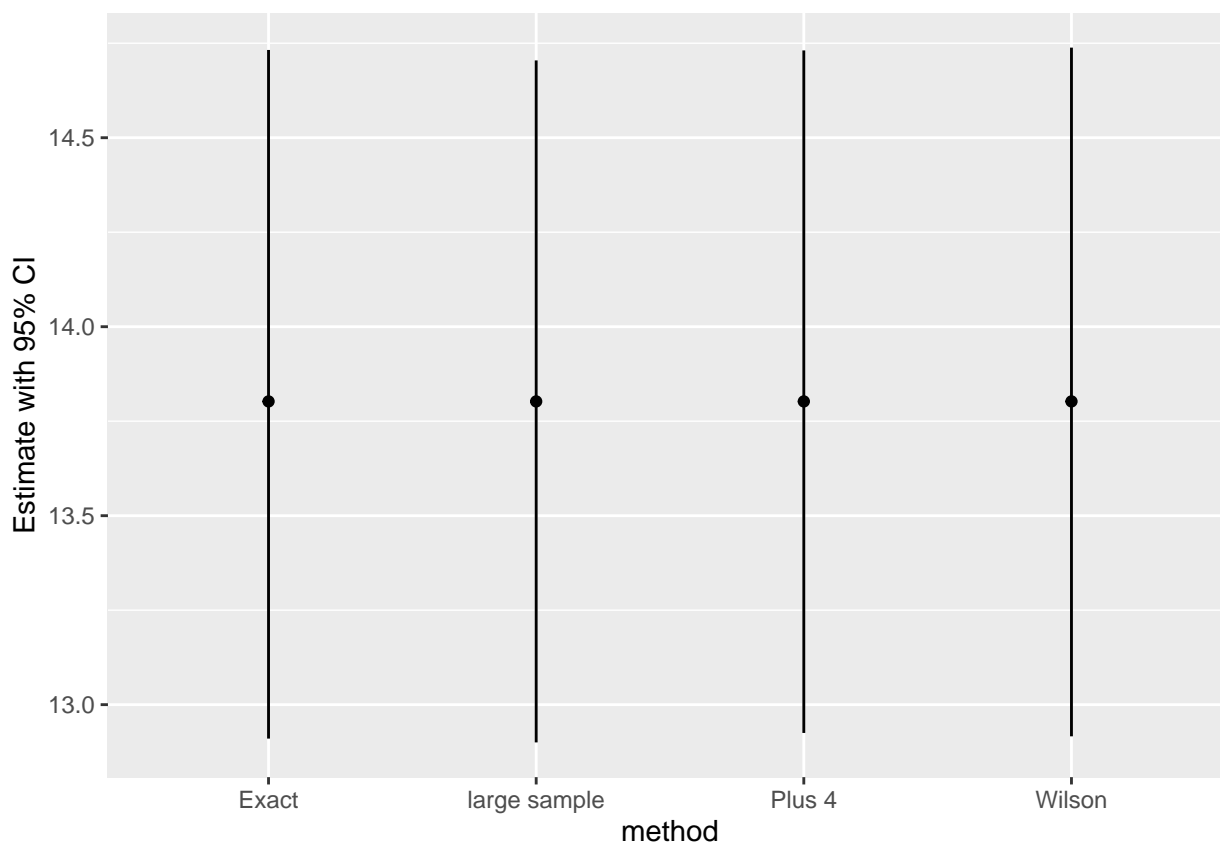
```
##
```

```
## All tests passed!
```

8. Create a plot to compare the confidence interval bounds. If you are stuck, refer back to the example code presented in lab 10.

```
breastfeed_CIs <- tibble(method = c("large sample", "Exact", "Wilson", "Plus 4"),
  lower_CI = c(12.90011, 12.91020, 12.91619, 12.92517),
  upper_CI = c(14.70452, 14.73222, 14.73842, 14.73099),
  estimate = c(p_hat*100, p_hat*100, p_hat*100, p_hat*100)
)

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
```



**9. Do the methods produce confidence intervals that are basically the same or very different? Why?**

The plot comparing the 4 CIs is below. They are nearly identical. This is because the sample size is large enough such that the CLT holds, implying that the large sample method is good, and so are all the other methods. When the sample size is large enough, all the CIs should agree.



**10. 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% falls far above all of the CIs. Because 18.6 is outside of the range of the CIs, we can conclude that the p-value for the corresponding hypothesis test would be  $< 5\%$  and conclude that yes, there is evidence in favor of the alternative hypothesis that the rate differs from 18.6%

To confirm your answer to question 10, perform a two-sided hypothesis test and interpret the p-value.

**11. State the null and alternative hypotheses.**

$H_0 : \mu = 18.6\%$   $H_a : \mu \neq 18.6\%$

12. Calculate the test statistic.

```
n <- 5615
p0 <- 0.186

p12 <- ((p_hat - p0) / sqrt(p0 * (1-p0) / n))
p12
```

```
## [1] -9.239275
```

```
. = ottr::check("tests/p12.R")
```

```
##
```

```
## All tests passed!
```

z-test for one-sample test for a proportion:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.1380232 - 0.186}{\sqrt{\frac{0.186(1-0.186)}{5615}}} = -9.239266$$

The test statistic is equal to -9.239266.

13. Calculate the p-value.

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

```
## [1] 2.48172e-20
```

```
. = ottr::check("tests/p13.R")
```

```
##
```

```
## All tests passed!
```

**14. Interpret this p-value in the context of the question.**

The p-value is very very tiny, much less than 0.0001%. This implies that the chance of seeing a proportion of 13.8% (or one even more different in magnitude) from the null value of 18.6% is  $< 0.0001\%$ . Thus, there is evidence against the null hypothesis, in favor of the alternative hypothesis that the proportion differs from 18.6%.

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 years old). Before 2007, no girls received the vaccine, while in the 4 years after it was introduced nearly 40% of girls received the vaccine each year. One concern 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.

For the following questions, you only need to read the abstract of the paper. 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.

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).

**15. 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 vs. girls who were not vaccinated.

**16. 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 infections?**

Given that the null value for the  $H_0$  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. Estimate the difference between the two proportions.**

```
succ1 <- 18
n1 <- 51

succ2 <- 5
n2 <- 47

p17 <- (18/51) - (5/47) # 35.3% - 10.6%
p17
```

```
## [1] 0.2465582
```

```
. = ottr::check("tests/p17.R")
```

```
##
## All tests passed!
```



18. 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 the object p18.

```
p1_tilde <- (succ1 + 1)/(n1 + 2)
p2_tilde <- (succ2 + 1)/(n2 + 2)
se <- sqrt((p1_tilde*(1 - p1_tilde)/(n1 + 2)) + (p2_tilde*(1 - p2_tilde)/(n2 + 2)))

p18 <- c((p1_tilde - p2_tilde) - 2.576 * se, (p1_tilde - p2_tilde) + 2.576 * se)
p18
```

```
## [1] 0.02784538 0.44423779
```

```
. = ottr::check("tests/p18.R")
```

```
##
```

```
## All tests passed!
```

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

Because the number of “successes” was 5 in the group who consumed peanuts regularly. Since  $5 < 10$ , it is not appropriate to use the large sample method.

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. State the null and alternative hypotheses.**

$$H_0 : p_1 = p_2 \quad H_A : p_1 \neq p_2$$

21. Calculate the test statistic.

```
phat <- (succ1 + succ2)/(n1 + n2)
p21 <- (succ1/n1 - succ2/n2)/sqrt(phat*(1-phat)*(1/n1 + 1/n2))
p21
```

```
## [1] 2.877213
```

```
. = ottr::check("tests/p21.R")
```

```
##
```

```
## All tests passed!
```

First, calculate  $\hat{p}$ , the estimated probability of having a peanut allergy assuming that the proportions are the same:  $\hat{p} = \frac{18+5}{51+47} = 0.2346939$

Then, the test statistic is: 
$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.3529412 - 0.106383}{\sqrt{0.2346939(1-0.2346939)\left(\frac{1}{51} + \frac{1}{47}\right)}} = 2.877213$$

**22. Calculate the p-value.**

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

```
## [1] 0.004012052
```

```
. = ottr::check("tests/p22.R")
```

```
##
```

```
## All tests passed!
```

**23. Interpret the p-value in the context of this question.**

The p-value is  $< 0.001$ . Because the p-value is so small there is evidence against the null hypothesis in favor of the alternative that there is a difference between the groups.

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

```
p24 <- '-1.20 to 0.90'
```

```
. = ottr::check("tests/p24.R")
```

```
##
```

```
## All tests passed!
```

25. 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  $\mu_d$ . Based on the p-value which of the following are true?

```
p25 <- 'A 95% CI will not contain 0, but a 99% CI will.'
```

```
. = ottr::check("tests/p25.R")
```

```
##
```

```
## All tests passed!
```

END