Assignment 8

Your name and student ID

Today's date

- Due date: November 8, 5:00pm (make sure to provide enough time for Gradescope submission to be uploaded if you include large image files).
- Late penalty: 50% late penalty if submitted within 24 hours of due date, no marks for assignments submitted thereafter.
- Submission process: Please submit a PDF of your assignment to Gradescope. You must tell Gradescope which questions are on which pages. If you can't see it properly on Gradescope, open the PDF in a PDF viewer at the same time so you can make the selections accurately, otherwise points will be deducted since this makes grading much less efficient.

Helpful hint:

- Knit your file early and often to minimize knitting errors! If you copy and paste code, you are bound to get an error that is hard to diagnose. Hand-writing 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!

- 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. When it doesn't, 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.

Question 1 [3 points total]. 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 5 percentage points.

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

<Write your answer here.>

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

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%.

1.b) [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 $p^* = 0.092$.

<Write your answer here.>

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

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%.

1.c) [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?

<Write your answer here.>

Solution: 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.

Question 2 [12 points total]. 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.

2.a) [4 points]. 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.

```
library(tidyverse)
```

```
#your code here.

#Solution:

#large sample
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
c(p_hat - 1.96*se, p_hat + 1.96*se) # CI
```

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

```
#exact
binom.test(num_successes, sample_size, p=0.5)
```

```
##
##  Exact binomial test
##
## data:  num_successes and sample_size
## 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
```

```
#wilson score
prop.test(num_successes, sample_size, p=0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  num_successes out of sample_size, 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
```

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

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

Summarize your findings:

- 95% CI using ________ method:
- 95% CI using ________ method:
- 95% CI using ________ method:
- 95% CI using ________ method:

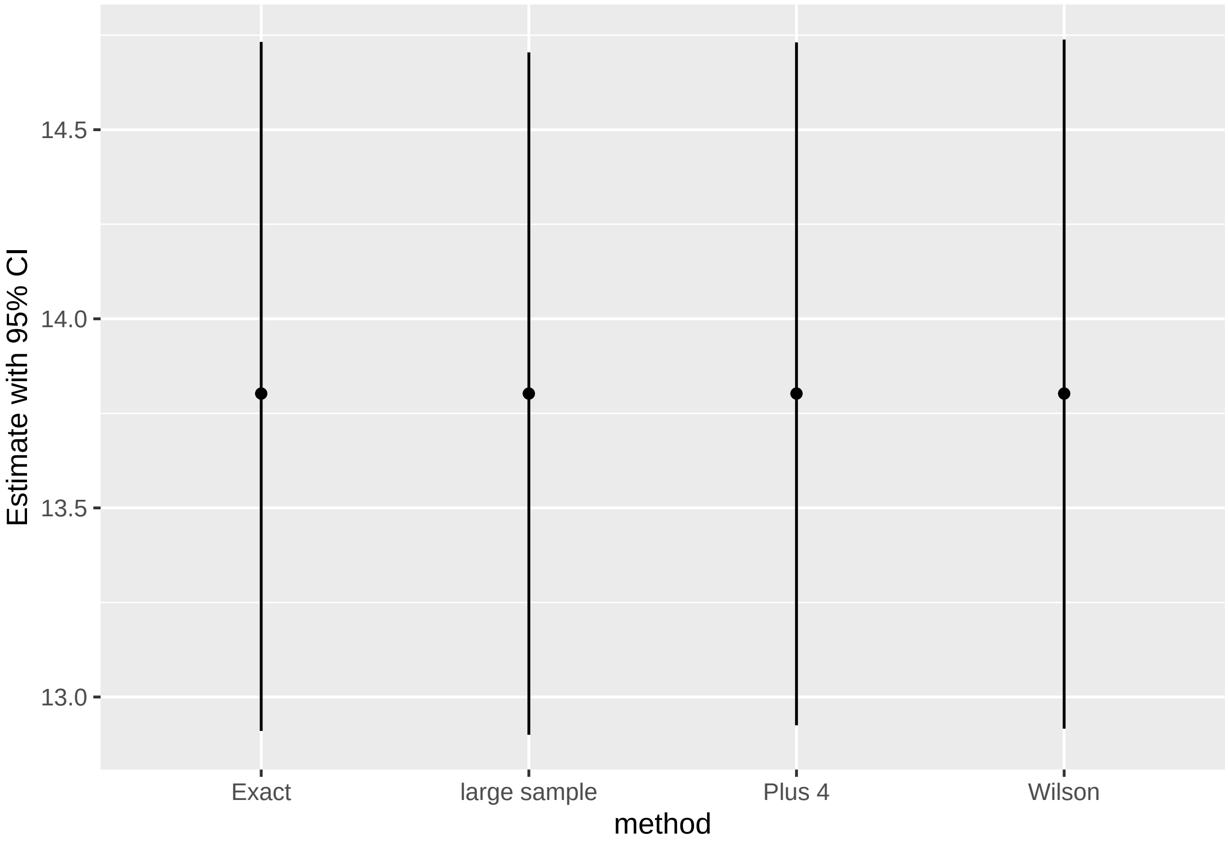
2. b) [3 points] Create a plot comparing the confidence intervals. Do the methods produce confidence intervals that are basically the same or very different? Why?

<Write your answer here.>

Solution: 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.

```
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)
                  )

ggplot(data = 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")
```



2.c) [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 2.b) 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%?

<Write your answer here.>

Solution: 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%

2.d) [4 points] To confirm your answer to part c), perform a two-sided hypothesis test and interpret the p-value.

State the null and alternative hypotheses:

<Write your answer here.>

Calculate the test statistic:

<Write your answer here. You may or may not want to insert a code chunk.>

Calculate the p-value:

```
#your code here.
```

Interpret the p-value:

<Write your answer here.>

Solution:

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

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

$$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.1380232-.186}{\sqrt{\frac{.186(1-.186)}{5615}}} = -9.239266$$

The test statistic is equal to -9.239266.

Calculate the p-value

```r
pnorm(-9.239266, lower.tail = T)*2
```

```
## [1] 2.481939e-20
```

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%.

Question 3 [2 points total]. 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.

3.a) 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)

[1 point] In particular, what does -0.61 estimate?

<Write your answer here.>

Solution: -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.

3.b) [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?

<Write your answer here.>

Solution: 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%.

Question 4 [7 points total]. An allergy to peanuts is increasingly comment in Western countries. A randomized controlled trial enrolled infants with a diagnosed peanut sensitivity. Infants were randomized to either avoide peanuts or consume them regularly until they reached age 5. At the end of the study 18 out of the 51 randomized to avoidance were allergic to peanuts compared to 5 out of the 47 randomized to consuming them regularly.

4.a) [1 point] Estimate the difference between the two proportions.

<Write your answer here. You can insert an R chunk to perform calculations if you wish.>

```r
succ1 <- 18
n1 <- 51

succ2 <- 5
n2 <- 47

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

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

4.b) [1 point] Use the plus four method to find a 99% confidence interval for the difference between the two groups.

<Write your answer here. You can insert an R chunk to perform calculations if you wish.>

Solution:

```r
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)) )

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

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

The 99% confidence interval for the difference is 2.78% to 44.4%. .

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

<Write your answer here.>

Solution: 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.

4.d) [4 points] 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.

<Write your answer here.>

Solution:

$H_0 : p_1 = p_2$ vs. $H_0 : p_1 \neq p_2$

Test statistic:

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: $\frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.3529412 - 0.106383}{\sqrt{0.2346939(1-0.2346939)\left(\frac{1}{51}+\frac{1}{47}\right)}} = 2.877213$

```r
pnorm(q = 2.877213, lower.tail = F)*2
```

```
## [1] 0.004012047
```

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.

5. [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?

a. -2.30 to -0.70
b. -1.20 to 0.90
c. 1.50 to 3.80
d. 4.50 to 6.90

Solution: b

6. [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?

a. Both a 95% CI and a 99% CI will contain 0.
b. A 95% CI will contain 0, but a 99% CI will not.
c. A 95% CI will not contain 0, but a 99% CI will.
d. Neither a 95% CI nor a 99% CI interval will contain 0.

Solution: c