Problem Set 8

name and student ID

Today's date

```
library(testthat)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:testthat':
##
##     matches
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.7     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::edition_get()   masks testthat::edition_get()
## x dplyr::filter()        masks stats::filter()
## x purrr::is_null()       masks testthat::is_null()
## x dplyr::lag()           masks stats::lag()
## x readr::local_edition() masks testthat::local_edition()
## x tidyr::matches()       masks dplyr::matches(), testthat::matches()
```

```
library(tibble)
```

Instructions

- Solutions will be released on Friday, November 4th.
- This semester, problem sets 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!
- To avoid code running off the page, 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.

Section 1: High school e-cigarette usage

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 to use. Suppose that you would like the width of the 95% confidence interval to be 2.5 percentage points.

1. Determine the most conservative sample size required to create a confidence interval of width 2.5 percentage points and assign it to the object p1. Recall that to do this, you need to use a p^* of 0.5.

```
. = " # BEGIN PROMPT
p1 <- NULL # YOUR CODE HERE
p1
" # END PROMPT

# BEGIN SOLUTION
p1 <- ceiling((1.96/0.0125)^2*0.5*(1-0.5))
# END SOLUTION
```

```
test_that("p1a", {
  expect_true(all.equal(p1, 6147, tol = 0.01))
  print("Checking: Sample size is correct")
})
```

```
## [1] "Checking: Sample size is correct"
## Test passed
```

$n = (z*/m)^2 p*(1-p*)$ $n = (1.96/0.0125)^2 \times 0.5 \times (1-0.5) = 6146.56 = 6147$ Thus, we would need a sample size of 6147 high school students to obtain a margin of error of 1.25 percentage points (width = 2.5 percentage points) if we assume the true prevalence is 50%.

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

```
. = " # BEGIN PROMPT
p2 <- NULL # YOUR CODE HERE
p2
" # END PROMPT


# BEGIN SOLUTION
p2 <- ceiling((1.96/0.0125)^2*0.092*(1-0.092))
# END SOLUTION
```

```
test_that("p2a", {
  expect_true(all.equal(p2, 2054, tol = 0.01))
  print("Checking: Sample size is correct")
})
```

```
## [1] "Checking: Sample size is correct"
## Test passed
```

$n = (z*/m)^2 p*(1-p*)$ $n = (1.96/0.0125)^2 \times 0.092 \times (1-0.092) = 2053.836 = 2054$ Thus, we would need a sample size of 2054 high school students to obtain a margin of error of 1.25 percentage points (width = 2.5 percentage points) if we assume the true prevalence is 9.2%.

3. The recent publication referenced in the previous question only looked at adults (aged 18+), but observed that the rate of current e-cigarette use was inversely related to age among the population they surveyed. Because of this finding would you suppose that the sample size estimated in question 2 is too low or too high? Assign your letter choice ("a" or "b") to `p3`.

a) too low
b) too high

```
. = " # BEGIN PROMPT
p3 <- NULL # YOUR ANSWER CHOICE HERE
p3
" # END PROMPT


# BEGIN SOLUTION
p3 <- "a"
# END SOLUTION
```

```
test_that("p3a", {
  expect_true(p3 == "a")
  print("Checking: Correct answer choice")
})
```

```
## [1] "Checking: Correct answer choice"
## Test passed
```

I would suppose that the estimated sample size is too low because the true prevalence among high school 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.

Section 2: Breastfeeding

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.

Use all four of the methods discussed in lecture and lab to create a 95% confidence interval for the proportion of California infants who are exclusively breastfed.

Use the concatenate function, `c()` to store your lower and upper bounds.

4. Use the large sample method of constructing a 95% CI. Do not round the lower or upper bounds.

```
. = " # BEGIN PROMPT
p4 <- NULL # YOUR CODE HERE
p4
" # END PROMPT


# BEGIN SOLUTION
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
# END SOLUTION
```

```
test_that("p4a", {
  expect_true(all.equal(p4[1], 0.1290011, 0.001))
  print("Checking: Correct lower bound")
})
```

```
## [1] "Checking: Correct lower bound"
## Test passed
```

```
test_that("p4b", {
  expect_true(all.equal(p4[2], 0.1470452, 0.001))
  print("Checking: Correct upper bound")
})
```

```
## [1] "Checking: Correct upper bound"
## Test passed
```

5. Use the Clopper Pearson (Exact) method for constructing a 95% CI. Do not round the lower or upper bounds.

```
. = " # BEGIN PROMPT
p5 <- NULL # YOUR CODE HERE
p5
" # END PROMPT
```

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

```
test_that("p5a", {
  expect_true(all.equal(p5[1], 0.1291020, 0.001))
  print("Checking: Correct lower bound")
})
```

```
## [1] "Checking: Correct lower bound"
## Test passed
```

```
test_that("p5b", {
  expect_true(all.equal(p5[2], 0.1473222, 0.001))
  print("Checking: Correct upper bound")
})
```

```
## [1] "Checking: Correct upper bound"
## Test passed
```

6. Use the Wilson Score method of constructing a 95% CI with a continuity correction. Do not round the lower or upper bounds.

```
. = " # BEGIN PROMPT
p6 <- NULL # YOUR CODE HERE
p6
" # END PROMPT


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

```
test_that("p6a", {
  expect_true(all.equal(p6[1], 0.1291619, 0.001))
  print("Checking: Correct lower bound")
})
```

```
## [1] "Checking: Correct lower bound"
## Test passed
```

```
test_that("p6b", {
  expect_true(all.equal(p6[2], 0.1473842, 0.001))
  print("Checking: Correct upper bound")
})
```

```
## [1] "Checking: Correct upper bound"
## Test passed
```

7. Use the Plus Four method of constructing a 95% CI. Do not round the lower or upper bounds.

```
. = " # BEGIN PROMPT
p7 <- NULL # YOUR CODE HERE
p7
" # END PROMPT

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

```
test_that("p7a", {
  expect_true(all.equal(p7[1], 0.1292549, 0.001))
  print("Checking: Correct lower bound")
})
```

```
## [1] "Checking: Correct lower bound"
## Test passed
```

```
test_that("p7b", {
  expect_true(all.equal(p7[2], 0.1473067, 0.001))
  print("Checking: Correct upper bound")
})
```

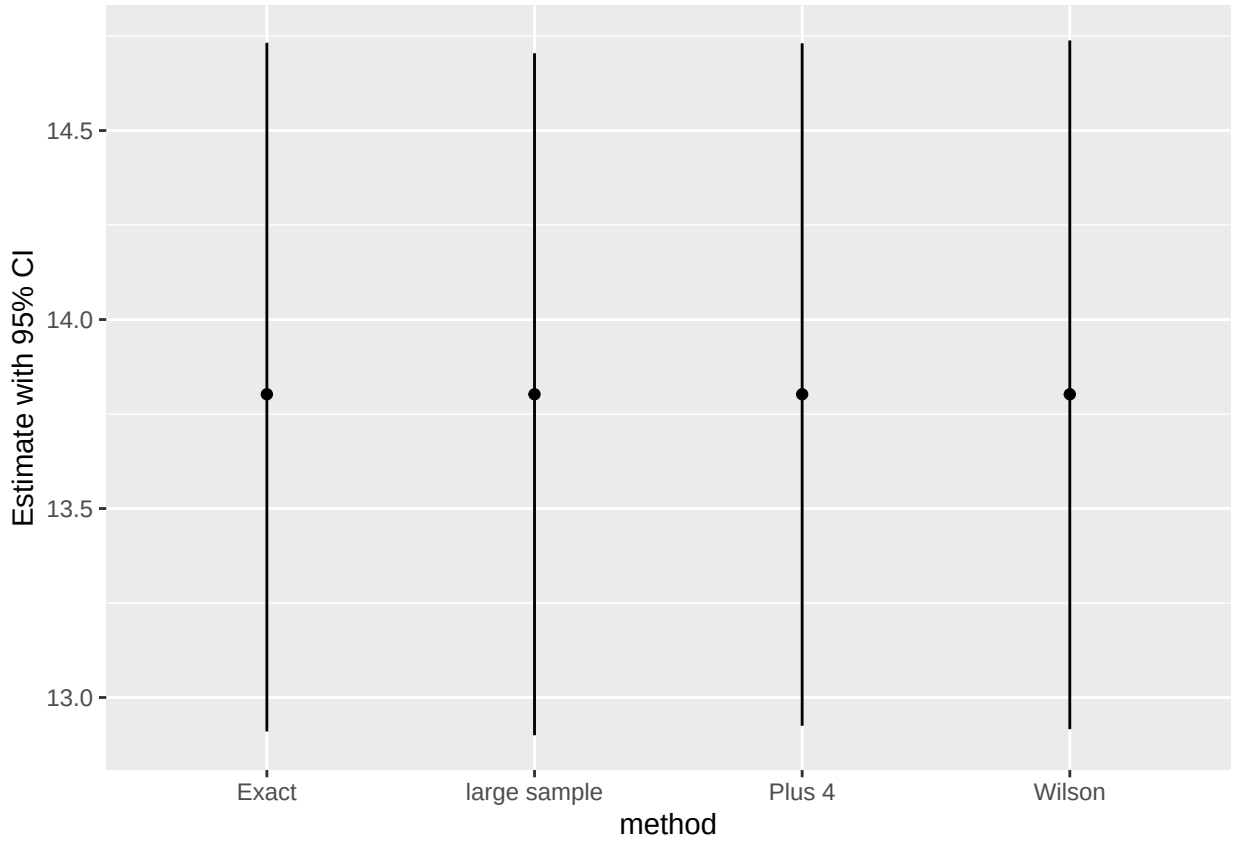
```
## [1] "Checking: Correct upper bound"
## Test passed
```

8. Create a table called `breastfeed_CIs` that contains each of the methods in the rows and their corresponding confidence interval lower and upper bounds in the columns. Then create a plot comparing the differences in confidence intervals by each method. If you are stuck, refer back to the example code presented in Lab 8.

```
. = " # BEGIN PROMPT
p8 <- NULL # YOUR CODE HERE
p8
" # END PROMPT

# BEGIN SOLUTION
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
```




```
# END SOLUTION
```

```
test_that("p8a", {
  expect_true("ggplot" %in% class(p8))
  print("Checking: p8 is a ggplot")
})
```

```
## [1] "Checking: p8 is a ggplot"
## Test passed
```

```
test_that("p8b", {
  expect_true(identical(p8$data, breastfeed_CIs))
  print("Checking: Used the correct data")
})
```

```
## [1] "Checking: Used the correct data"
## Test passed
```

```
test_that("p8c", {
  expect_true(rlang::quo_get_expr(p8$mapping$x) == "method")
  print("Checking: Method is on the x-axis")
})
```

```
## [1] "Checking: Method is on the x-axis"
## Test passed
```

```
test_that("p8d", {
  expect_true(rlang::quo_get_expr(p8$mapping$y) == "estimate")
  print("Checking: Estimate is plotted as a point")
})
```

```
## [1] "Checking: Estimate is plotted as a point"
## Test passed
```

```
test_that("p8e", {
  expect_true('GeomSegment' %in% sapply(p8$layers, function(x) class(x$geom)[1]))
  print("Checking: Made line segments of confidence intervals")
})
```

```
## [1] "Checking: Made line segments of confidence intervals"
## Test passed
```

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

The plot comparing the 4 CIs shows that the intervals are nearly identical. This is because the sample size is large enough such that the CLT holds, implying that the large sample method provides an accurate estimate of the interval, and so do all of 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. Do not round your answer.

```
. = " # BEGIN PROMPT
p12 <- NULL # YOUR CODE HERE
p12
" # END PROMPT

# BEGIN SOLUTION
n <- 5615
p0 <- 0.186

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

# END SOLUTION
```

```
test_that("p12a", {
  expect_true(all.equal(p12, -9.239275, 0.001))
  print("Checking: correct test statistic")
})
```

```
## [1] "Checking: correct test statistic"
## Test passed
```

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

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

The test statistic is equal to -9.239266.

13. Calculate the p-value. Do not round your answer.

```
. = " # BEGIN PROMPT
p13 <- NULL # YOUR CODE HERE
p13
" # END PROMPT

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

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

```
# END SOLUTION
```

```
test_that("p13a", {
  expect_true(all.equal(p13, 2.48172e-20, 0.001))
  print("Checking: Correct p-value")
})
```

```
## [1] "Checking: Correct p-value"
## Test passed
```

14. Interpret the p-value in the context of this 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%.

Section 3: HPV Vaccine

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 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. 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). 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 diseases?

Given that the null value 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%.

Section 4: Peanut Allergy

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.

```
. = " # BEGIN PROMPT
p17 <- NULL # YOUR CODE HERE
p17
" # END PROMPT

# BEGIN SOLUTION
succ1 <- 18
n1 <- 51

succ2 <- 5
n2 <- 47

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

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

```
# END SOLUTION
```

```
test_that("p17a", {
  expect_true(all.equal(p17, 0.2465582, 0.001))
  print("Checking: Correct estimated difference")
})
```

```
## [1] "Checking: Correct estimated difference"
## Test 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 an object called p18.

```
. = " # BEGIN PROMPT
p18 <- NULL # YOUR CODE HERE
p18
" # END PROMPT

# BEGIN SOLUTION
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)

# END SOLUTION
```

```
test_that("p18a", {
  expect_true(all.equal(p18[1], 0.02784538, 0.001))
  print("Checking: Correct lower bound")
})
```

```
## [1] "Checking: Correct lower bound"
## Test passed
```

```
test_that("p18b", {
  expect_true(all.equal(p18[2], 0.44423779, 0.001))
  print("Checking: Correct upper bound")
})
```

```
## [1] "Checking: Correct upper bound"
## Test passed
```

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

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.

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

21. Calculate the test statistic.

```
. = " # BEGIN PROMPT
p21 <- NULL # YOUR CODE HERE
p21
" # END PROMPT

# BEGIN SOLUTION
phat <- (succ1 + succ2)/(n1 + n2)
p21 <- (succ1/n1 - succ2/n2)/sqrt(phat*(1- phat)*(1/n1 + 1/n2))

# END SOLUTION
```

```
test_that("p21a", {
  expect_true(all.equal(p21, 2.877213, 0.001))
  print("Checking: Correct test statistic")
})
```

```
## [1] "Checking: Correct test statistic"
## Test 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.

```
. = " # BEGIN PROMPT
p22 <- NULL # YOUR CODE HERE
p22
" # END PROMPT

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

```
test_that("p22a", {
  expect_true(all.equal(p22, 0.004012052, 0.001))
  print("Checking: Correct p-value")
})
```

```
## [1] "Checking: Correct p-value"
## Test passed
```

23. Interpret the p-value in the context of this problem.

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 obtain a p-value of 0.21. Which one of the following could be a possible 95% confidence interval for μ_d? Assign the letter of your answer choice to `p24`. For example, `p24 <- "a"`.

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

```
. = " # BEGIN PROMPT
p24 <- NULL # YOUR ANSWER CHOICE HERE
p24
" # END PROMPT

# BEGIN SOLUTION
p24 <- "b"
# END SOLUTION
```

```
test_that("p24a", {
  expect_true(p24 == "b")
  print("Checking: Correct answer choice")
})
```

```
## [1] "Checking: Correct answer choice"
## Test 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 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.

```
. = " # BEGIN PROMPT
p25 <- NULL # YOUR ANSWER CHOICE HERE
p25
" # END PROMPT

# BEGIN SOLUTION
p25 <- "c"
# END SOLUTION
```

```
test_that("p25a", {
  expect_true(p25 == "c")
  print("Checking: Correct answer choice")
})
```

```
## [1] "Checking: Correct answer choice"
## Test passed
```