Lab 09: Proportions and Inference

- Due date: Friday, November 6 at 10:00pm.
- 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.

Instructions

Tests of changes in sex ratios based on a single sample

There is a long literature studying changes in sex-ratios of births due to stressful events, such as 9/11. In today's lab, we consider a relatively small study that recorded biomarkers of stress on pregnancy. In the group of subjects that had the highest markers of stress (based on cortisol), there were 14 births to males out of a total of 38.

In this lab, we will compare the four methods we learned to calculate CIs for proportions. Recall that two of these methods involved hand calculations (though we can treat R as if it were a calculator) and two of the methods used built-in R functions.

1. Use the Normal approximation to construct a 95% confidence interval in this high stress group. We also called this specific method of constructing the CI the "large sample method". Assign the object `large_sample_ci` to a vector of the lowerbound and upperbound rounded to 4 decimals

```r
###############################################

# example of final answer where 0.1234 is my
# lowerbound and 0.5678 is my upperbound
my_ci <- c(0.1234, 0.5678)

##############################################

p.hat <- 14/38
se <- sqrt(p.hat * (1 - p.hat)/38)
p.hat - 1.96 * se
```

```
## [1] 0.2150476
```

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

```
## [1] 0.5217945
```

```r
large_sample_ci <- c(0.2150, 0.5218)

large_sample_ci
```

```
## [1] 0.2150 0.5218
```

```r
check_problem1()
```

```
## [1] "Checkpoint 1 Passed: Answer is numeric."
## [1] "Checkpoint 2 Passed: Upperbound and lowerbound have been stored."
## [1] "Checkpoint 3 Passed: Correct, the lowerbound is 0.2150476 and upperbound 0.5217945."
##
## Problem 1
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

2. Create the 95% CI again, this time using the R function that implements the Wilson Score method with a continuity correction.

```r
prop.test(x = 14, n = 38, conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  14 out of 38, null probability 0.5
## X-squared = 2.1316, df = 1, p-value = 0.1443
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2229295 0.5400424
## sample estimates:
##         p
## 0.3684211
```

```r
wilson_score_CI <- c(0.2229, 0.5400)


wilson_score_CI
```

```
## [1] 0.2229 0.5400
```

```r
check_problem2()
```

```
## [1] "Checkpoint 1 Passed: Answer is numeric."
## [1] "Checkpoint 2 Passed: Upperbound and lowerbound have been stored."
## [1] "Checkpoint 3 Passed: Correct, the lowerbound is 0.2229 and upperbound 0.5400."
##
## Problem 2
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

3. Create the 95% CI again, this time using the Plus 4 method.

```r
p.tilde <- (14 + 2)/(38 + 4)
se <- sqrt(p.tilde * (1 - p.tilde)/42)
p.tilde - 1.96 * se
```

```
## [1] 0.2340838
```

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

```
## [1] 0.5278209
```

```r
plus_4_ci <- c(0.2341, 0.5278)


plus_4_ci
```

```
## [1] 0.2341 0.5278
```

```r
check_problem3()
```

```
## [1] "Checkpoint 1 Passed: Answer is numeric."
## [1] "Checkpoint 2 Passed: Upperbound and lowerbound have been stored."
## [1] "Checkpoint 3 Passed: Correct, the lowerbound is 0.2340838 and upperbound 0.5278209."
##
## Problem 3
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

4. Create the 95% CI again, this time using the R function that implements the Clopper Pearson (Exact) method.

```r
binom.test(x = 14, n = 38, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  14 and 38
## number of successes = 14, number of trials = 38, p-value = 0.1433
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2181250 0.5400572
## sample estimates:
## probability of success
##              0.3684211
```

```r
exact_method_ci <- c(0.2181, 0.5401)
exact_method_ci
```

```
## [1] 0.2181 0.5401
```

```r
check_problem4()
```

```
## [1] "Checkpoint 1 Passed: Answer is numeric."
## [1] "Checkpoint 2 Passed: Upperbound and lowerbound have been stored."
## [1] "Checkpoint 3 Passed: Correct, the lowerbound is 0.2181 and upperbound 0.5401."
##
## Problem 4
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

5. Summarize the four methods' estimates in the following table. Do they include the null hypothesized value for the sex ratio?

Method	95% Confidence Interval
Large sample	21.50% to 52.18%
Wilson Score*	22.29% to 54.00%
Plus four	23.41% to 52.78%
Exact	21.81% to 54.01%

- with continuity correction.

The null hypothesized value of the sex ratio is 14/38 = 36.84%. This value is included in all four of the calculated confidence intervals.

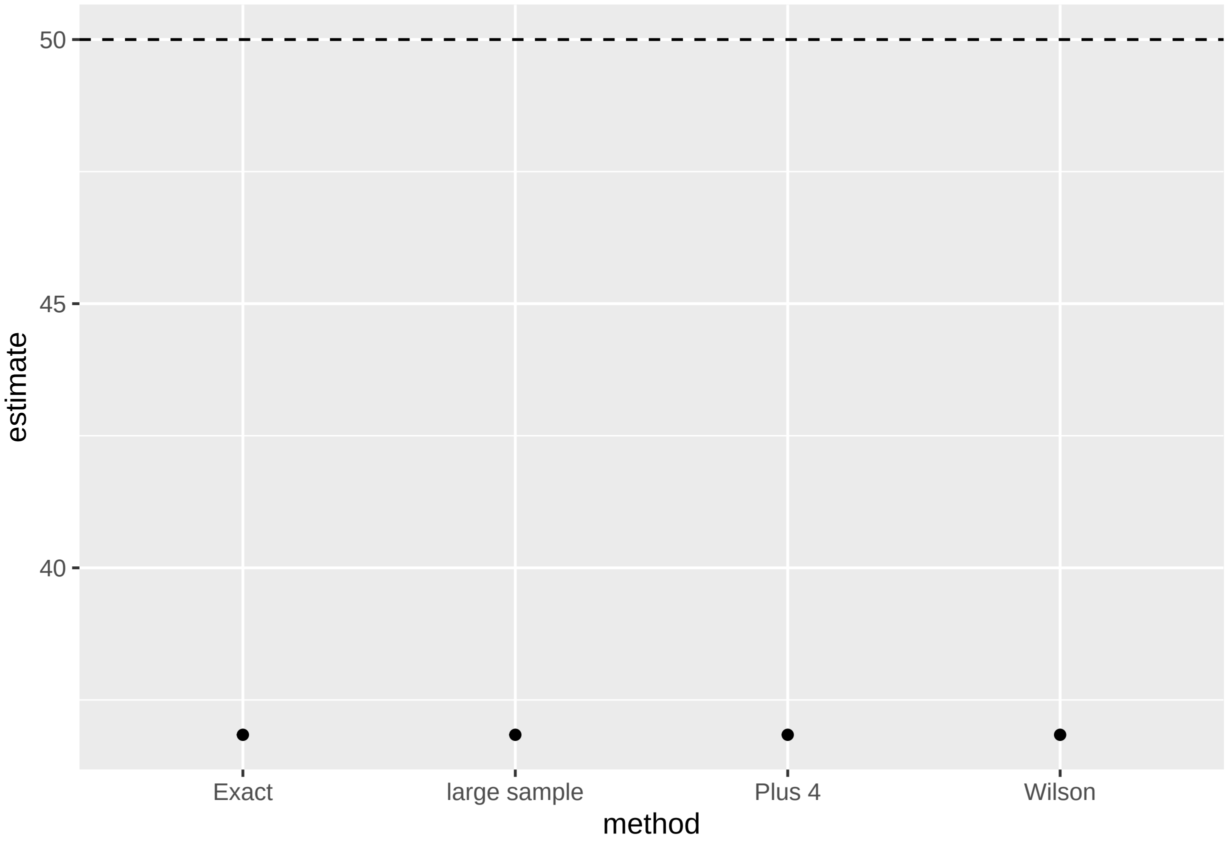
6. Here is a graphical representation of the confidence intervals.

```r
# First make a tibble (an easy way to make a data frame) with the data about
# each confidence interval. To do this, replace each instance of 0.00 with the
# estimate from your calculations above.
library(ggplot2)
library(tibble)
```

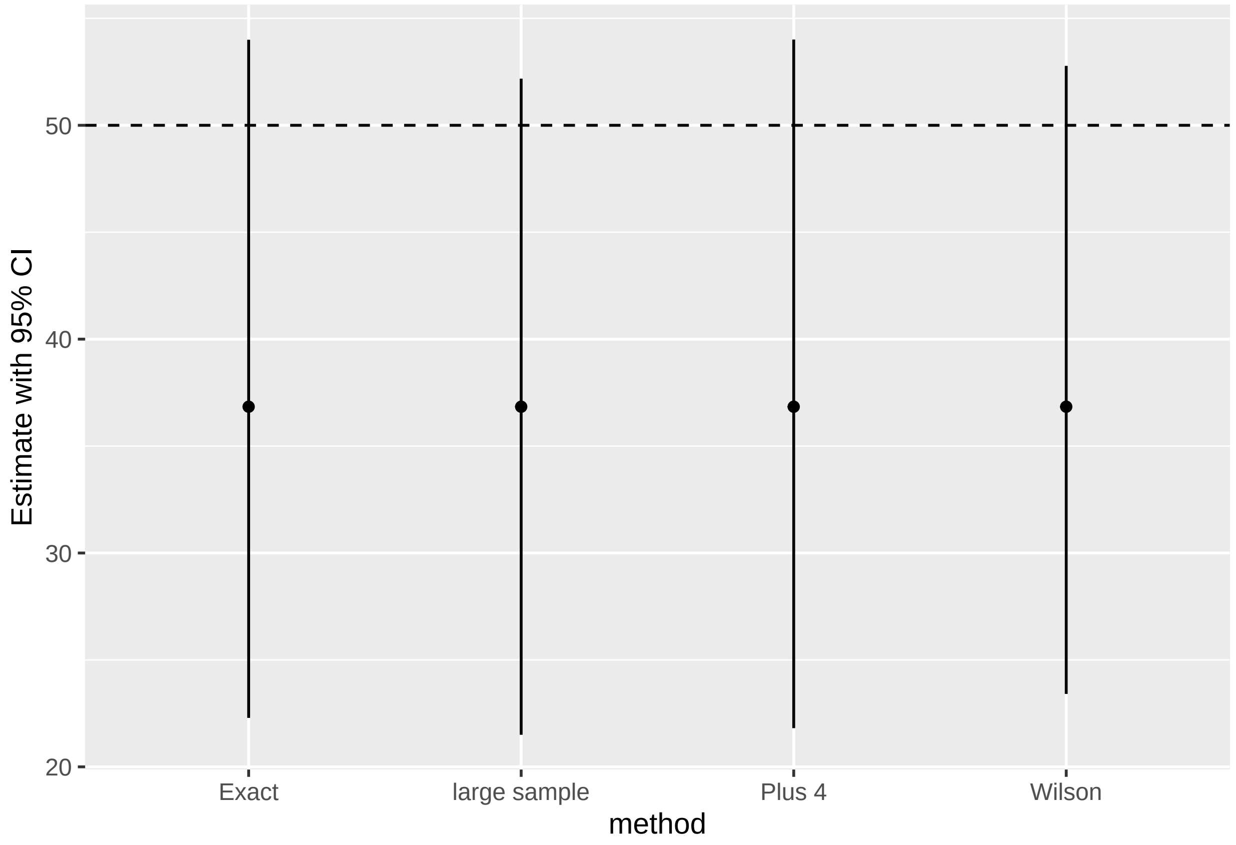
```
##
## Attaching package: 'tibble'

## The following object is masked from 'package:assertthat':
##
##     has_name
```

```r
sex_CIs <- tibble(method   = c("large sample", "Exact", "Wilson", "Plus 4"),
                  lower_CI = c(21.50          , 22.29    , 23.41    , 21.81),
                  upper_CI = c(52.18          , 54.00    , 52.78    , 54.01),
                  estimate = c(36.84          , 36.84    , 36.84    , 36.84)
                  )
# Build the ggplot incrementally, to understand how it works.
# Step 1: (qu: why do we put a horizontal line at 50?)
ggplot(data = sex_CIs, aes(x = method, y = estimate)) +
  geom_point() +
  geom_hline(aes(yintercept = 50), lty = 2)
```

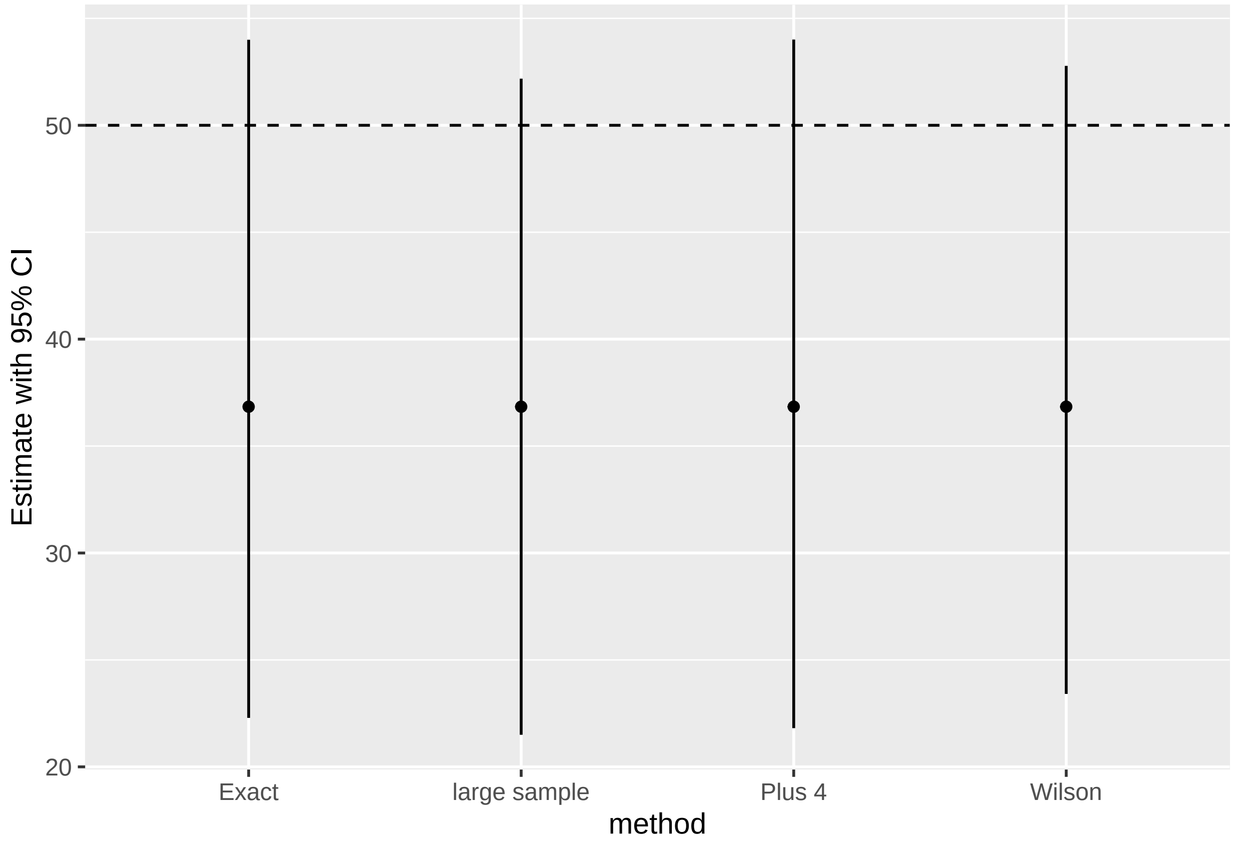


```
# Step 2:
ggplot(data = sex_CIs, aes(x = method, y = estimate)) +
  geom_point() +
  geom_hline(aes(yintercept = 50), lty = 2) +
  geom_segment(aes(x = method, xend = method, y = lower_CI, yend = upper_CI)) +
  labs(y = "Estimate with 95% CI")
```



```
p6 <- ggplot(data = sex_CIs, aes(x = method, y = estimate)) +
  geom_point() +
  geom_hline(aes(yintercept = 50), lty = 2) +
  geom_segment(aes(x = method, xend = method, y = lower_CI, yend = upper_CI)) +
  labs(y = "Estimate with 95% CI")


p6
```



```
check_problem6()
```

```
## [1] "Checkpoint 1 Passed: A ggplot has been defined."
## [1] "Checkpoint 2 Passed: sex_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."
##
## Problem 6
## Checkpoints Passed: 4
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

What does `geom_segment()` do? In particular, what do `x`, `xend`, `y` and `yend` specify in this case?

geom_segment() shows us our confidence intervals as vertical lines. x is the starting point of the line and xend is end, both of which in this case have to do with the individual methods on the x-axis. y gives us the lower bound of our interval (starting point of the vertical line) and yend gives us the upper bound of our interval (ending point of the vertical line).

7. Based on this plot, what can you say about the confidence intervals for the sex ratio in the high stress group?

The confidence intervals calculated by the 4 different methods are overall pretty similar and close. However, the interval calculated by the Wilson method is slightly smaller compared to the other confidence intervals. The confidence intervals also all capture the true proportion of 14/38.

8. If you have time, repeat the above analysis for the group with low stress. There were 25 births to this group, of which 17 of them were to males.

```
# large sample
p.hat <- 17/25
se <- sqrt(p.hat * (1 - p.hat)/25)
p.hat - 1.96 * se
```

```
## [1] 0.4971413
```

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

```
## [1] 0.8628587
```

```
c(0.4971413, 0.8628587)
```

```
## [1] 0.4971413 0.8628587
```

```
# wilson
prop.test(x = 17, n = 25, conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  17 out of 25, null probability 0.5
## X-squared = 2.56, df = 1, p-value = 0.1096
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4644983 0.8427132
## sample estimates:
##    p
## 0.68
```

```
c(0.4644983, 0.8427132)
```

```
## [1] 0.4644983 0.8427132
```

```
# plus four
p.tilde <- (17 + 2)/(25 + 4)
se <- sqrt(p.tilde * (1 - p.tilde)/29)
p.tilde - 1.96 * se
```

```
## [1] 0.4821765
```

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

```
## [1] 0.8281683
```

```r
c(0.4821765, 0.8281683)
```

```
## [1] 0.4821765 0.8281683
```

```r
# exact
binom.test(x = 17, n = 25, conf.level = 0.95)
```

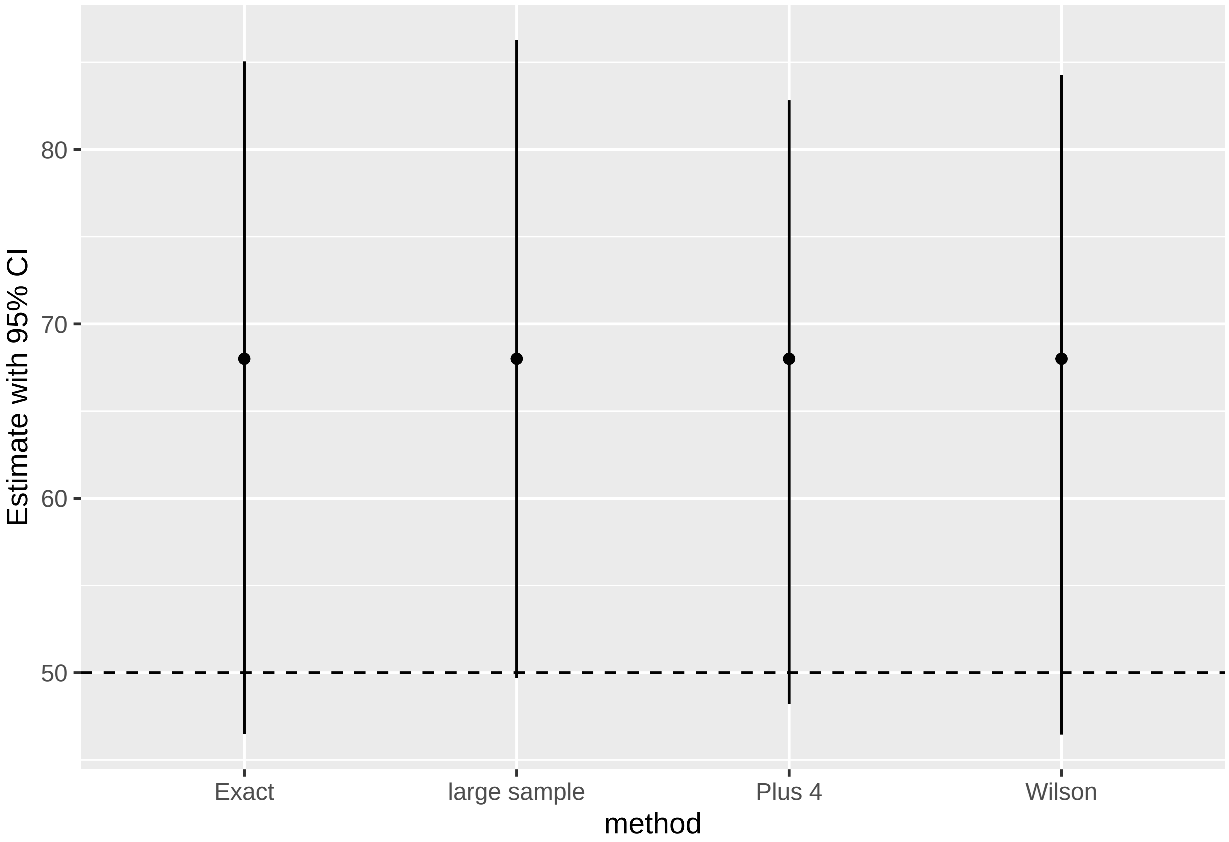
```
##
##  Exact binomial test
##
## data:  17 and 25
## number of successes = 17, number of trials = 25, p-value = 0.1078
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4649993 0.8505046
## sample estimates:
## probability of success
##                   0.68
```

```r
c(0.4649993, 0.8505046)
```

```
## [1] 0.4649993 0.8505046
```

```r
library(ggplot2)
library(tibble)
sex_CIs_2 <- tibble(method   = c("large sample", "Exact", "Wilson", "Plus 4"),
                  lower_CI = c(49.71           , 46.50    ,  46.455   ,  48.22),
                  upper_CI = c(86.29           , 85.05    ,  84.27    ,  82.82),
                  estimate = c(68.00           , 68.00    ,  68.00    ,  68.00)
                  )
p8 <- ggplot(data = sex_CIs_2, aes(x = method, y = estimate)) +
  geom_point() +
  geom_hline(aes(yintercept = 50), lty = 2) +
  geom_segment(aes(x = method, xend = method, y = lower_CI, yend = upper_CI)) +
  labs(y = "Estimate with 95% CI")

p8
```



9. If you recreated the graph for the low stress group, what can you say about the confidence intervals for the sex ratio in this group?

It is more variable compared to the last group of confidence intervals, and also contains higher values.

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         PASSED               1    autograded
## Problem 3         PASSED               1    autograded
## Problem 4         PASSED               1    autograded
## Problem 5 NOT YET GRADED               1 free-response
## Problem 6         PASSED               1    autograded
## Problem 7 NOT YET GRADED               1 free-response
## Problem 8 NOT YET GRADED               1 free-response
## Problem 9 NOT YET GRADED               1 free-response
```

Submission

For assignments in this class, you'll be submitting using the **Terminal** tab in the pane below. In order for the submission to work properly, make sure that:

1. Any image files you add that are needed to knit the file are in the `src` folder and file paths are specified accordingly.
2. You **have not changed the file name** of the assignment.
3. The file is saved (the file name in the tab should be **black**, not red with an asterisk).
4. The file knits properly.

Once you have checked these items, you can proceed to submit your assignment.

1. Click on the **Terminal** tab in the pane below.
2. Copy-paste the following line of code into the terminal and press enter.

```
cd; cd ph142-fa20/lab/lab09; python3 turn_in.py
```

3. Follow the prompts to enter your Gradescope username and password. When entering your password, you won't see anything come up on the screen–don't worry! This is just for security purposes–just keep typing and hit enter.
4. If the submission is successful, you should see "Submission successful!" appear as output.
5. If the submission fails, try to diagnose the issue using the error messages–if you have problems, post on Piazza.

The late policy will be strictly enforced, **no matter the reason**, including submission issues, so be sure to submit early enough to have time to diagnose issues if problems arise.