W4: Synchronous class

This activity uses the data we collected using the Zoom poll in the Week 3 synchronous class.

Assuming none of you can 'read' windings, there was no information that should have prompted respondents to prefer one answer option over the others.

Therefore, shouldn't we expect that respondents are equally likely to pick each option and any differences in proportions are due to chance acting alone? Well, there is a known effect where respondents to multi-choice questions tend prefer answer option C when they are just guessing. This activity will explore this phenomenon with YOUR data for the windings question.

For this investigation we are interested in exploring if respondents are picking answer option C a different proportion of the time than we'd expect just by chance.

Load the tidyverse and the data

```
library(tidyverse)
winding_answers = tibble(Answer=c(rep('A', 15+41),
                                  rep('B', 24+32),
                                  rep('C', 35+48),
                                  rep('D', 14+23)),
                         sesh=c(rep('Morning', 15), rep('Afternoon',41),
                                rep('Morning', 24), rep('Afternoon',32),
                                rep('Morning', 35), rep('Afternoon',48),
                                rep('Morning', 14), rep('Afternoon',23))) %>%
  rowid_to_column()
write_csv(winding_answers, "windings_question.csv")
windings <- read_csv("windings_question.csv")
windings %>% glimpse()
```

```
## Rows: 232
## Columns: 3
## $ rowid  <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
## $ Answer <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A"~
## $ sesh   <chr> "Morning", "Morning", "Morning", "Morning", "Morning", "Morning~
```

Question 1

What are an appropriate null and alternative hypothesis for this test?

A. $H_0 : p_C = 0.5$ & $H_1 : p_C \neq 0.5$

B. $H_0 : \hat{p}_C = 0.25$ & $H_1 : \hat{p}_C \neq 0.25$

C. $H_0 : p_C = 0.25$ & $H_1 : p_C \neq 0.25$

D. $H_0 : \hat{p}_C = 0.5$ & $H_1 : \hat{p}_C \neq 0.5$

E. $H_0 = \hat{p}_C = 0.5$ & $H_1 = \hat{p}_C \neq 0.5$

F. $H_0 = p_C = 0.25$ & $H_1 = p_C \neq 0.25$

Question 2

What does the test statistic for this hypothesis test represent?

A. The proportion of students who picked option C in our class.

B. The proportion of students who picked each option.

C. The difference between the proportion of students who picked C and the proportion who picked other options.

D. The proportion of all students like the students in our class that would pick option C in a question like this.

Question 3

Modify this code (if needed) to calculate the test statistic for this hypothesis test.

```
windings %>%
  group_by(Answer) %>%
  summarise(n=n())
```

```
## # A tibble: 4 x 2
##   Answer     n
##   <chr>  <int>
## 1 A         56
## 2 B         56
## 3 C         83
## 4 D         37
```

A. 0.1595

B. 232

C. 3/4

D. 0.3578

E. 0.35

F. 0.25

Don't forget (1)

```
test_stat <- NA # ***replace NA with value here***
```

Once you have the correct answer from above, save it as test_stat, i.e. replace the NA.

Question 4

What value should you set `n_observations` equal to so that we can *simulate* the proportion of respondents picking C, assuming they are just picking at random? The choice needs to be appropriate for this investigation.

A. 1000

B. 232

C. 144

D. 88

Don't forget (2)

```
n_observations <- NA # ***add a value here***
```

Make sure you add the correct value above once you have determined it, i.e. replace the NA.

Question 5

Which of the following would be the best code to simulate the proportion of respondents who pick C, assuming **respondents pick at random**.

Tip: Check that probabilities and the sample size are appropriate for this investigation.

```
# A
a <- sample(c("C", "Not C"), size = n_observations, replace = FALSE)

# B
b <- sample(c("C", "Not C"), size = n_observations, prob = c(0.25, 0.75), replace = TRUE)

# C
c <- sample(c("A", "B", "C", "D"), prob = c(0.5, 0.5, 0.5, 0.5),
            size = n_observations, replace = FALSE)

# D
d <- sample(c("C"), prob = c(0.25), size = n_observations, replace = TRUE)
```

Question 6

Which of the following statements best describes the plot created by the code below?

A. The distribution of the proportion of respondents picking C in our class last week.

B. The estimated sampling distribution if respondents are not choosing their answer at random.

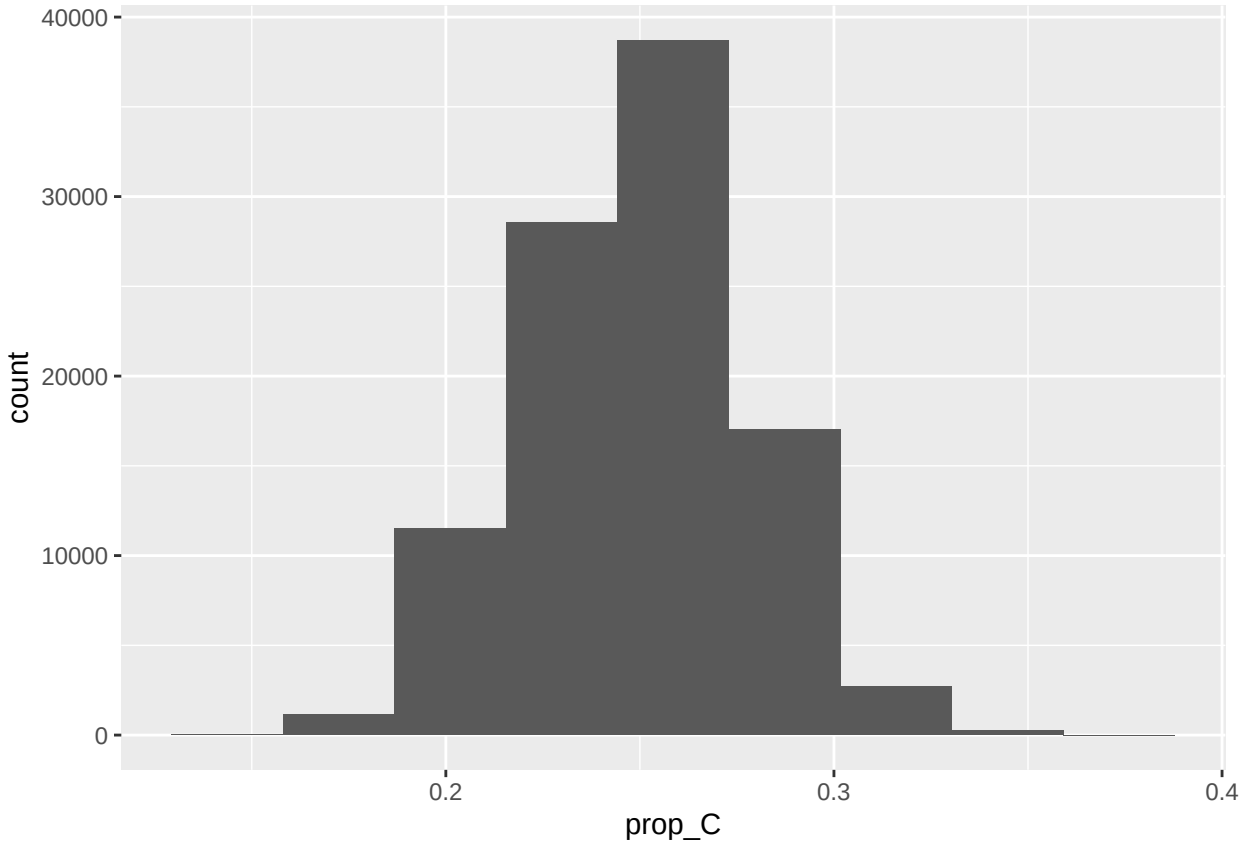
C. The estimated sampling distribution of the proportion of respondents picking option C if the null hypothesis is true.

D. The distribution we'd expect from repeated sampling of students in our class.

```
set.seed(2022)

repetitions <- 100000
simulated_stats <- rep(NA, repetitions)
# this vector will store the simulated proportions of respondents choosing C

for (i in 1:repetitions)
{
  new_sim <- sample(c("A", "B", "C", "D"), size=232, replace=TRUE)
  sim_p <- sum(new_sim == "C") / 232
  simulated_stats[i] <- sim_p
}
sim <- tibble(prop_C = simulated_stats)
ggplot(sim, aes(prop_C)) +
  geom_histogram(bins=10)
```

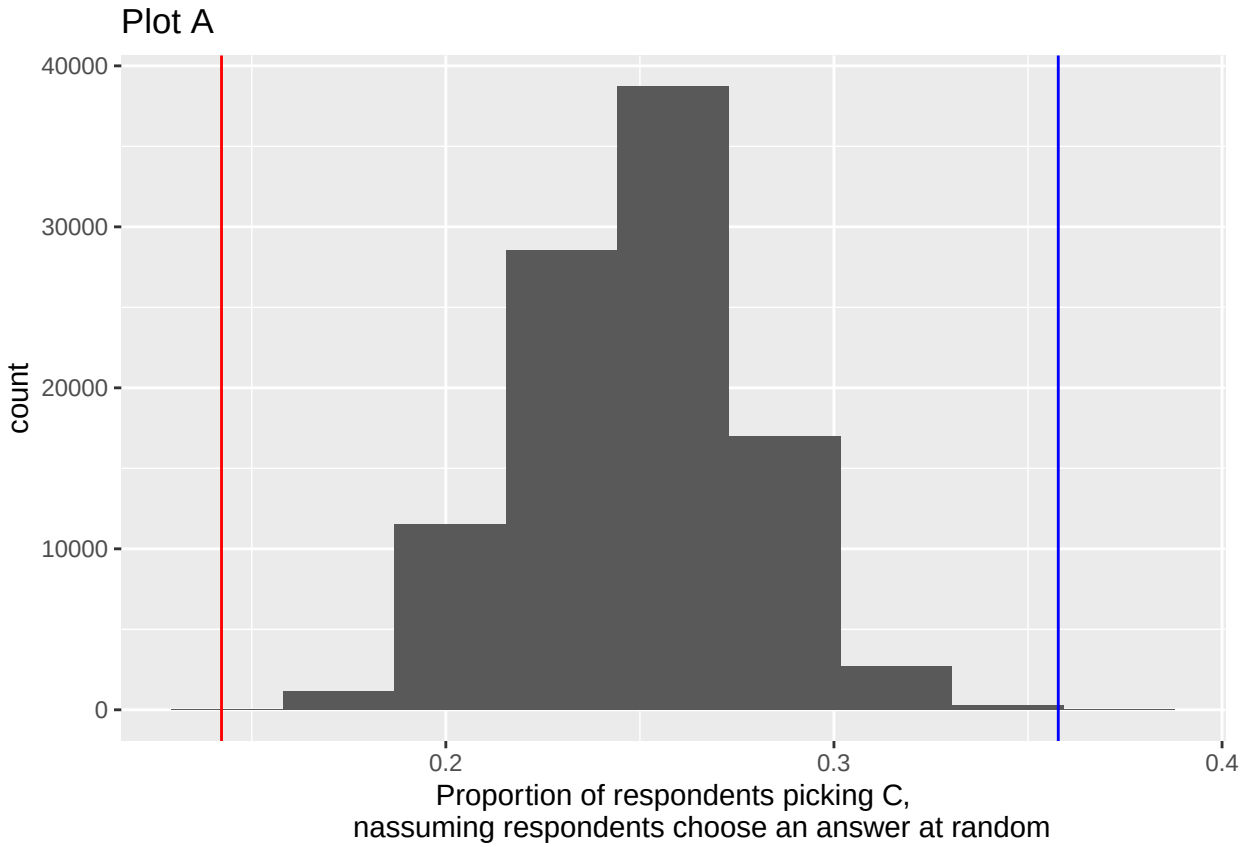


Question 7

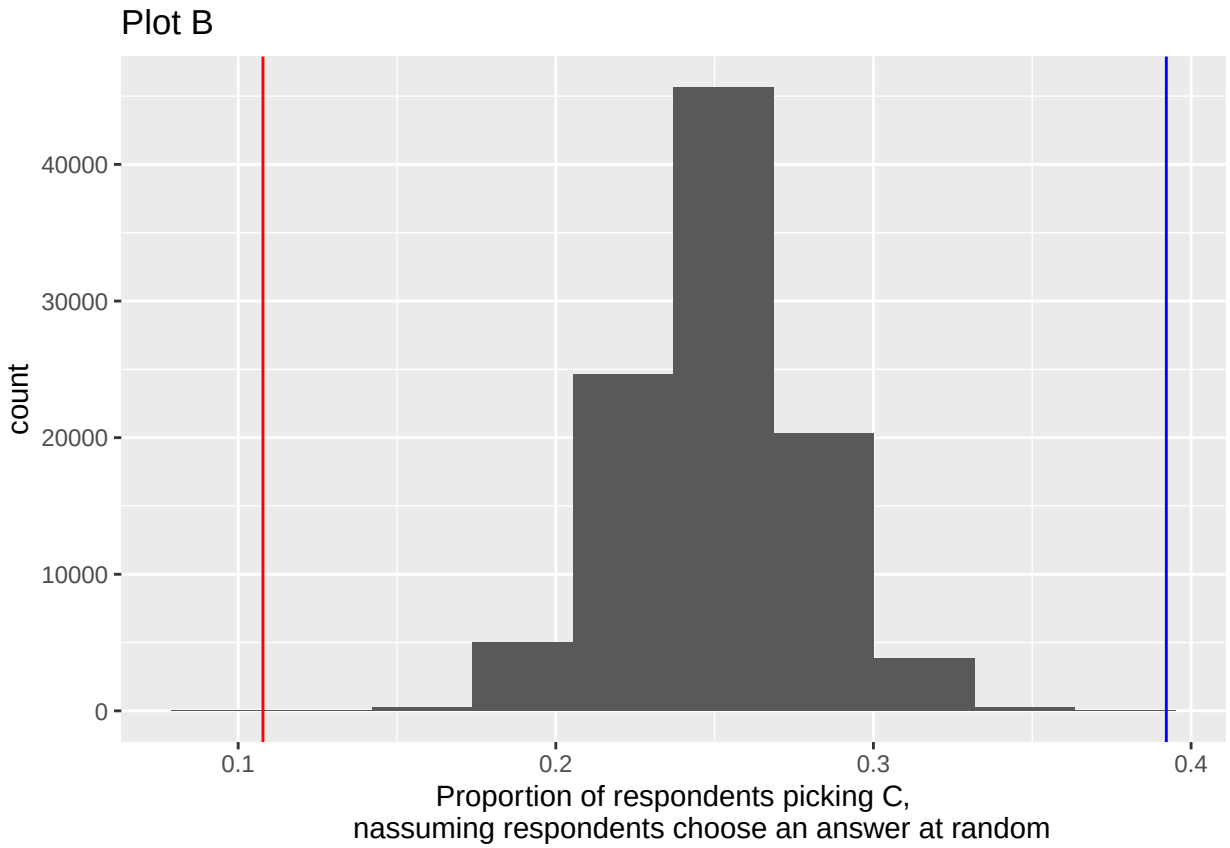
Which of the following graphs is appropriately marked for calculating the p-value for this hypothesis test?

```
base_plot <- ggplot(sim, aes(prop_C)) +
  geom_histogram(bins=10) +
  labs(x="Proportion of respondents picking C,\
       nassuming respondents choose an answer at random")
```

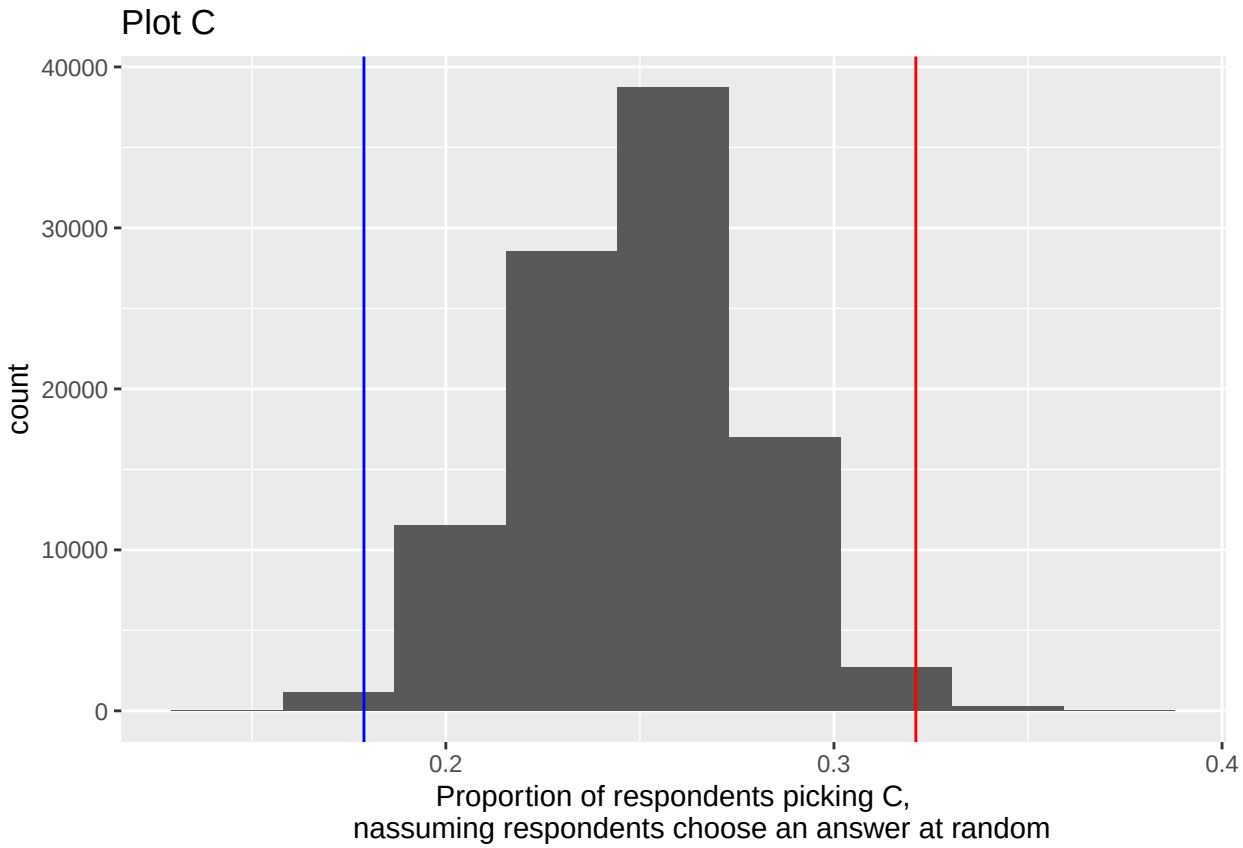
```
# A
base_plot +
  geom_vline(xintercept=0.3578, color="blue") +
  geom_vline(xintercept=0.25-(0.3578-0.25), color="red") +
  labs(title = "Plot A")
```



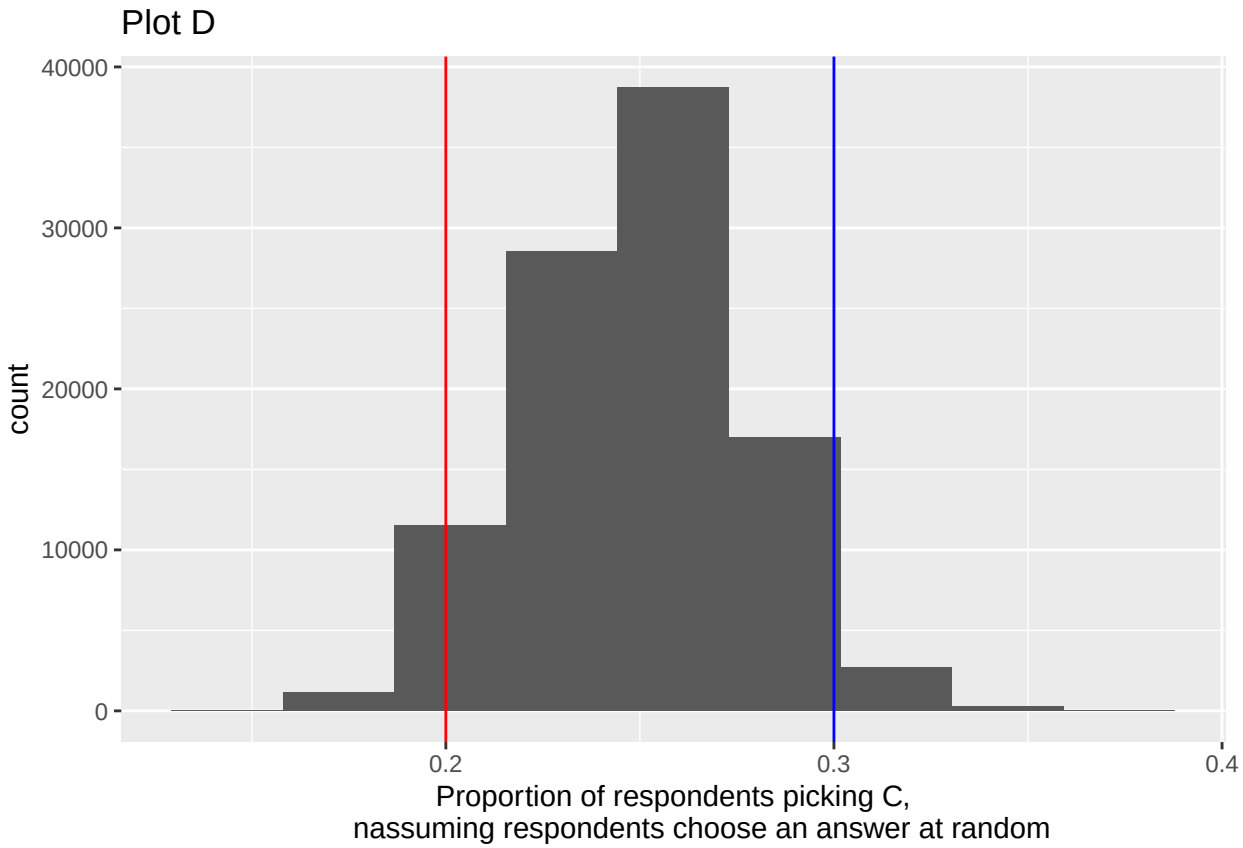
```r
# B
base_plot +
  geom_vline(xintercept=0.3578-0.25, color="red") +
  geom_vline(xintercept=0.25 + 0.25-(0.3578-0.25), color="blue") +
  labs(title = "Plot B")
```



```r
# C
base_plot +
  geom_vline(xintercept=0.25 + abs(0.3578-0.5)/2, color="red") +
  geom_vline(xintercept=0.25 + (0.3578-0.5)/2, color="blue") +
  labs(title = "Plot C")
```



```r
# D
base_plot +
  geom_vline(xintercept=0.2, color="red") +
  geom_vline(xintercept=0.3, color="blue") +
  labs(title = "Plot D")
```



Question 8

What is the p-value of this hypothesis test? Alter the following code (if needed).

```
sim %>%
  filter(prop_C >= test_stat | prop_C <= 0.25-(test_stat-0.25)) %>%
  summarise(p_value = n() / repetitions)
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

A. The p-value is below 0.0001.

B. The p-value is between 0.0001 and 0.01.

C. The p-value is between 0.01 and 0.05.

D. The p-value is between 0.05 and 0.1.

E. The p-value is above 0.1.

Question 9

Based off the p-value calculated in Question 8. What is the best interpretation of the p-value?

A. There is 0 evidence against the null hypothesis.

B. It is impossible to get a test statistic like ours or more extreme if the null hypothesis is true.

C. We have strong evidence in favour of the alternative hypothesis.

D. There is very strong evidence against the claim that students are picking their answers at random.

Question 10

Please select the option that attests that you worked with your group members on this.