Practice Problems

Question 1

Approximately 10% of the general population is left-handed. Suppose that the university is conducting a study to see if this percentage is the same among their students. This would help inform classroom renovations to ensure sufficient left-handed (and right-handed) seating. Suppose 500 students are randomly selected and asked whether or not they are left-handed. Suppose that 63 of these 500 students respond that they are left-handed.

(a) What are appropriate null and alternative hypothesis to test the claim?

The null hypothesis is: the porportion of university students who are left-handed is 0.10.

The alternative hypothesis is: the porportion of university students who are left-handed is not 0.10.

(b) Use the sample() function to simulate the number of left-handed students in a random sample of 500 students, assuming that the prevalence of left-handedness is the same among University of Toronto students as it is in the general population. How many left-handed students did you have in your simulated sample of 500 students? How does this simulated count compare to the results of the handedness study (i.e., that 63 of the 500 students sampled were left-handed)? How does it compare to the assumption that 10% of students are left-handed.

Note that the probabilities assigned to the values in the vector from which you're sampling using the sample() function are considered equal by default. For example, consider simulating flipping a coin 10 times:

```
sample(c("Head","Tail"),size=10,replace=TRUE)
```

```
##  [1] "Tail" "Tail" "Tail" "Head" "Tail" "Head" "Tail" "Head" "Tail" "Head"
```

```
# will do the same thing as:
sample(c("Head","Tail"),size=10, prob=c(0.5, 0.5), replace=TRUE)
```

```
##  [1] "Head" "Tail" "Head" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Head"
```

```
# Even though the exact counts of "Head" and "Tail" differ each time you
# run this code, if you simulate enough coin flips (by increasing
# the value of 'size', you'll get approximately the same proportion of "Head" and "Tail" outcomes)

# To modify the code to make Tails much more likely than Heads, we could change the probs:
sample(c("Head","Tail"),size=10,prob=c(0.2, 0.8), replace=TRUE)
```

```
##  [1] "Tail" "Tail" "Tail" "Head" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail"
```

Set the random number seed to the last digit of your student number before carrying out your simulation. We set the seed so that the results won't change each time this code is run or knitted. If we didn't do this, your interpretations and conclusions may not be relavant to the new run of the code (or when you knit your Rmd file!).

```
set.seed(6)
left_sim <- sample(c("Left_handed","Right-handed"), size=500, prob=c(0.1, 0.9), replace = TRUE)
total_left <- sum(left_sim == "Left_handed")
total_left
```

```
## [1] 61
```

There are 61 left-handed students (porportion: 61/500 = 0.122) in my simulated sample. This result is quite similar to the results of the handedness study (i.e., that 63 of the 500 students sampled were left-handed, porportion = 63/500 = 0.126). This simulated result is also quite similar to the assumption that 10% (0.1 porportion) of students are left-handed.

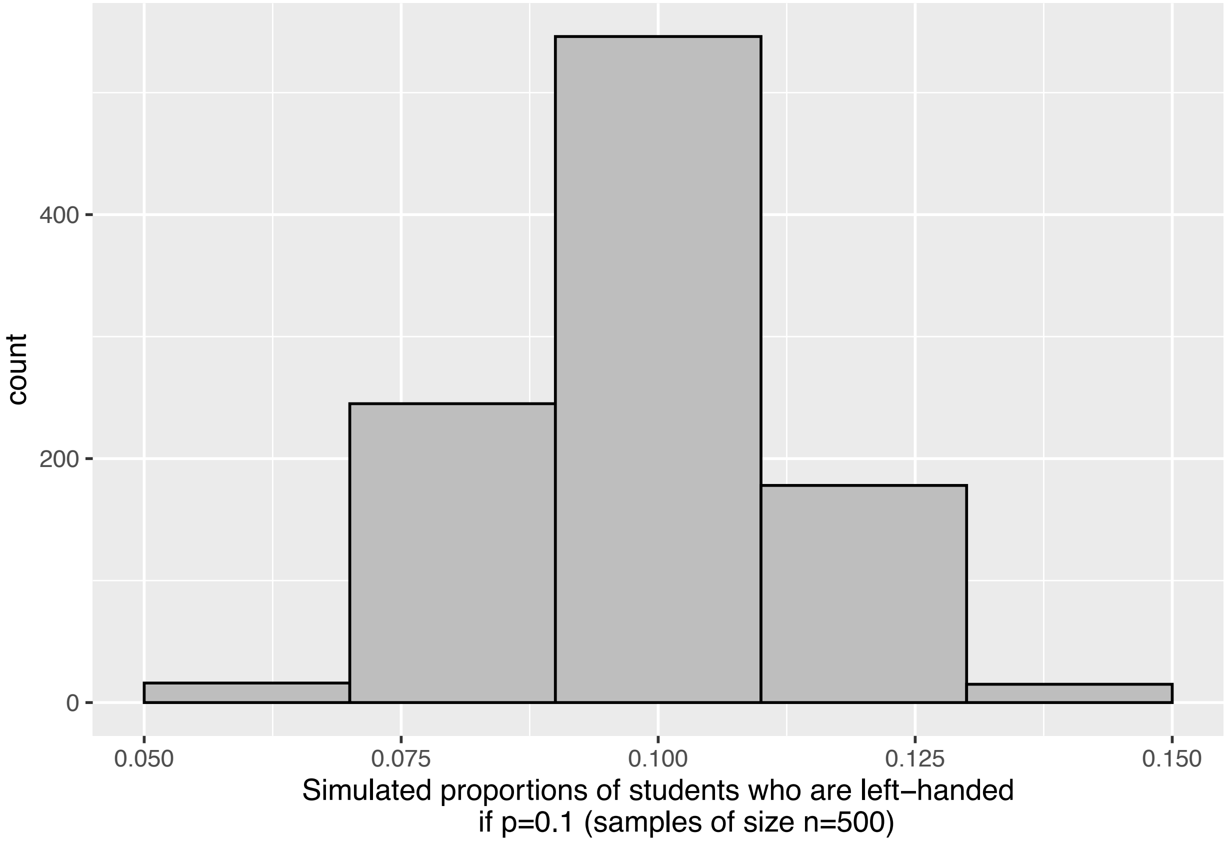
(c) Use R to estimate the sampling distribution of the test statistic under the assumption that the prevalence of left-handedness among University of Toronto students matches the general popuation. Use 1000 repetitions and set the seed to the last 2 digits of your student number. Generate the plot of this estimated sampling distribution and describe the distribution in a few sentences.

```
n_observations <- 500
repetitions <- 1000
simulated_stats <- rep(NA, repetitions)
set.seed(36)

#generate simulation
for (i in 1:repetitions){
left_sim <- sample(c("Left-handed", "Right-handed"),
                    size = n_observations,
                    prob = c(0.1,0.9),
                    replace = TRUE)
  left_p <- sum(left_sim == "Left-handed") / n_observations
  simulated_stats[i] <- left_p;
}

#save simulation
left <- data_frame(p_left_hand = simulated_stats)

#plot simulation
left %>% ggplot(aes(x = p_left_hand)) +
  geom_histogram(binwidth = 0.02, colour = "black", fill = "grey") +
  xlab("Simulated proportions of students who are left-handed
       if p=0.1 (samples of size n=500)")
```



The distribution of the simulated porportions is fairly symmetric and unimodal. It is centered at porportion = 0.100.

(d) Use R to compute the p-value of this hypothesis test based on the sampling distribution that you estimated in part (c).

```
left_pvalue <- left %>%
  filter(abs(p_left_hand - 0.1) >= abs(0.126 - 0.1)) %>%
  summarise(left_pvalue = n() / repetitions)
left_pvalue
```

```
## # A tibble: 1 x 1
##   left_pvalue
##         <dbl>
## 1       0.053
```

(e) Which of the following statements is/are valid description of the P-value you computed in (d):

i. The probability that the proportion of U of T students who are left-handed matches the general population.

ii. The probability that the proportion of U of T students who are left-handed does not match the general population.

iii. The probability of obtaining a number of left-handed students in a sample of 500 students at least as extreme as the result in this study.

iv. The probability of obtaining a number of left-handed students in a sample of 500 students at least as extreme as the result in this study, if the prevelance of left-handedness among all U of T students matches the general population.

i), iv)

(f) Write a conclusion to this hypothesis test based on the p-value you computed in part (d). Since p-value = 0.053 which is greater than 0.05 (assumed significance level). We do not reject the null hypothesis. There is weak evidence against the null hypothesis and we can conclude that the porportion of university students who are left-handed (0.10) does generally match the general population.

Question 2

A Scottish woman noticed that her husband's scent changed. Six years later he was diagnosed with Parkinson's disease. His wife joined a Parkinson's charity and noticed that odour from other people. She mentioned this to researchers who decided to test her abilities. They recruited 6 people with Parkinson's disease and 6 people without the disease. Each of the recruits wore a t-shirt for a day, and the woman was asked to smell the t-shirts (in random order) and determine which shirts were worn by someone with Parkinson's disease. She was correct for 11 of the 12 t-shirts! You can read about this here.

(a) Without conducting a simulation, describe what you would expect the sampling distribution of the proportion of correct guesses about the 12 shirts to look like if someone was just guessing.

The sampling distribution of the proportion of correct guesses about the 12 shirts to look like if someone was just guessing should be centered at porportion = 0.5, as there is a 50-50 percent chance of guessing right.

(b) Carry out a test using simulation to determine if there is evidence that this woman has some ability to identify Parkinson's disease by smell, or if she was a lucky guesser.

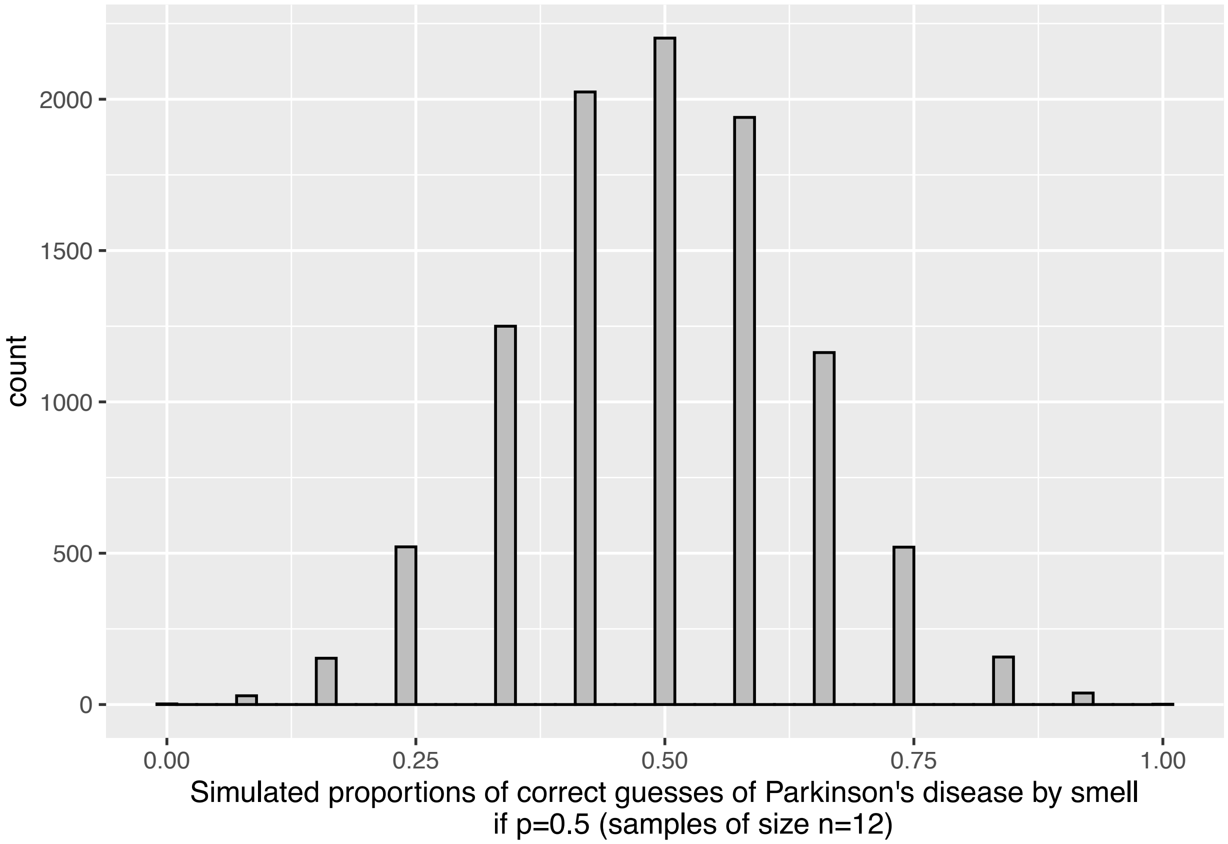
Set the random number seed to the last two digits of your student number before carrying out your simulation. Use 10,000 repetitions. (This simulation is similar to the code in Question 1, but with many more simulated values of the test statistic under the null hypothesis. 10,000 is a lot of repetitions - more that is likely needed - but we'll do this many repetitions this time anyways.

```r
n_observations <- 12
repetitions <- 10000
simulated_stats <- rep(NA, repetitions)
set.seed([redacted])

#generate simulation
for (i in 1:repetitions){
guess_sim <- sample(c("Correct", "Incorrect"),
                    size = n_observations,
                    prob = c(0.5,0.5),
                    replace = TRUE)
  guess_p <- sum(guess_sim == "Correct") / n_observations
  simulated_stats[i] <- guess_p;
}

#save simulation
guess <- data_frame(p_correct = simulated_stats)

#plot simulation
guess %>% ggplot(aes(x = p_correct)) +
  geom_histogram(binwidth = 0.02, colour = "black", fill = "grey") +
  xlab("Simulated proportions of correct guesses of Parkinson's disease by smell
       if p=0.5 (samples of size n=12)")
```



```r
#calculate p value
guess_pvalue <- guess %>%
  filter(abs(p_correct - 0.5) >= abs(11/12 - 0.5)) %>%
  summarise(guess_pvalue = n() / repetitions)
guess_pvalue
```

```
## # A tibble: 1 x 1
##   guess_pvalue
##          <dbl>
## 1        0.007
```

Since the P-value is 0.007 which is small than the assumed significance level (0.05) we conclude that we have we have strong evidence against the null hypothesis that the porportion of correct guesses for Parkinson's disease by smell from the woman is 0.5.

The data provide convincing evidence that this woman, in fact, she has abilities to guess Parkinson's disease by smell.

(c) The woman correctly identified all 6 people who had been diagnosed with Parkinson's but incorrectly identified one of the others as having Parkinson's. Eight months later he was was diagnosed with the disease. So the woman was actually correct 12 out of 12 times. Are you able to get the p-value for the test using this new data, without running a new simulation? What would you change from your answer to (b)? What wouldn't you change?

Yes, I am able to get the p-value for the test without running the simulation. The p-value would just be 0 instead of 0.007 since there are no possible data that has a porportion more extreme than 1. This means that their is clear evidence that the woman has abilities to guess Parinson's disease by smell.

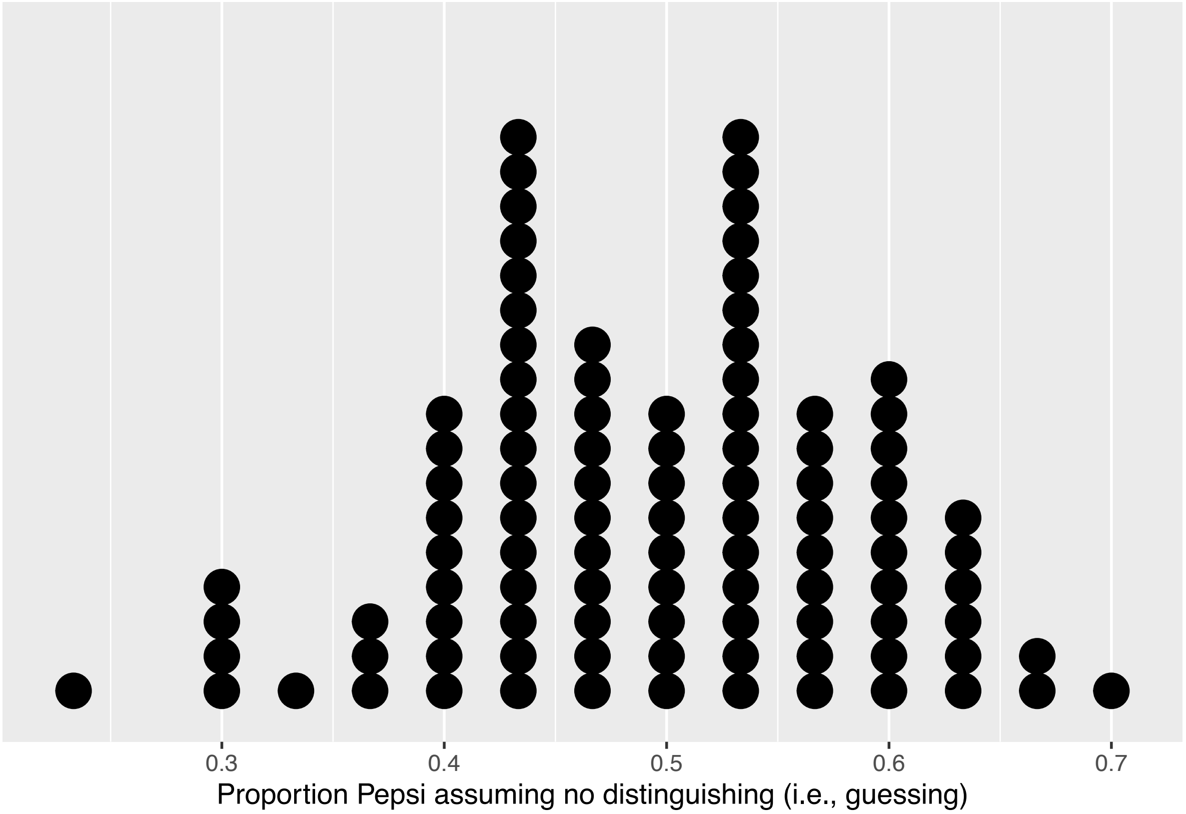
Question 3

(Adapted from ISRS 3.17) Some people claim that they can tell the difference between Coke or Pepsi in the first sip. In fact, pop drinkers often have strong preferences for one over the other. A researcher wanting to test the claim that people can distinguish (correctly or incorrectly) between them randomly selected 30 people. He then filled 30 plain white cups with Pepsi and asked each person to take one sip from their cup and identify the drink as Coke or Pepsi. 13 participants correctly identified the drink as Pepsi and the other 17 participants said the drink was Coke. Does this suggest that people can correctly (or incorrectly) tell a difference between the two drinks?

(a) What are appropriate null and alternative hypothesis to test the claim?

Null hypothesis: The porportion of correct guesses for Coke or Pepsi is 0.5. Alternative hypothesis: The porportion of correct guesses for Coke or Pepsi is not 0.5.

(b) Assume you conduct a test of significance using simulation and get the following estimated sampling distribution of the test statistic assuming the null hypothesis is true. For simplicity, this distribution shows the results of only **100** simulations. There are 100 dots on the plot, one for each simulation. (In practice, 100 simulations is not sufficient to obtain a good estimate the sampling distribution.)



(i)What does each single dot in the plot represent? One porportion value of correct guesses about Coke or Pepsi from one simulation.

(ii)Based on this plot, what is your estimate of the P-value?

```
pop_pvalue <- sim %>%
  filter(abs(p_correct - 0.5) >= abs(13/30 - 0.5)) %>%
  summarise(pop_pvalue = n() / repetitions)
pop_pvalue
```

```
## # A tibble: 1 x 1
##   pop_pvalue
##         <dbl>
## 1        0.63
```

(c) What conclusion can you make based on the p-value you calculated in part b(ii)? Since the p-value is greater than 0.05. There is no evidence against the null hypothesis. Therefore, the porportion of correct guesses for Coke or Pepsi is indeead approximately 0.5, which means that people do not have the ability to precisely tell the difference.

(d) Suppose the analysis described in (b) is repeated but this time 1000 simulations are used to get a better estimate of the P-value, and the resulting P-value was 0.04. What is an appropriate conclusion? Since the p-value is less than 0.05. We have moderate evidence against the null hypothesis and conclude that people are in fact able to tell the difference between Coke and Pepsi.

Question 4

Is C a more (or less) popular answer when students guess answers to multiple choice questions? Check out https://mhssoundtosea.com/1474/features/when-in-doubt-is-choice-c-really-the-best-answer/ for arguments why C might not be the option if you're just guessing!

Suppose a difficult PollEverywhere question (one on a topic that has not yet been discussed in class) with five options is asked during lecture. Students are instructed to answer on their own (i.e., no discussing with others) and to guess an answer if they don't know which one is correct. 34 of the 150 people who respond to the poll answer C. The rest of the students answer A, B, D or E.

(a) What are appropriate null and alternative hypotheses? Describe how you determined the null value? Null hypothesis: The porportion of students who chose C is 0.25. Alternative hypothesis: The porportion of students who chose C is not 0.25.

I determined the null value by assuming that students have no preference on the choices A,B,C,D, so there is an equally likely chance of choosing each choice. (1/4 = 0.25 chance for each choice)

(b) Sketch the sampling distribution you'd expect to see if people choose multiple choice answers randomly when guessing. *You do not need to include that sketch in your knitted pdf.* Write 2-3 sentences describing the distribution you sketched. Does it look similar to the estimated sampling distributions in previous questions in these practice problems? Why or why not?

The distribution of the porpotion of choosing C asssuming there is no preference for C is symmetrical and unimodal. It is centered at porportion = 0.25. The shape of the distribution looks similar to the other estimated sampling distributions in previous questions. However, the range for the data is less than the other sampling data.

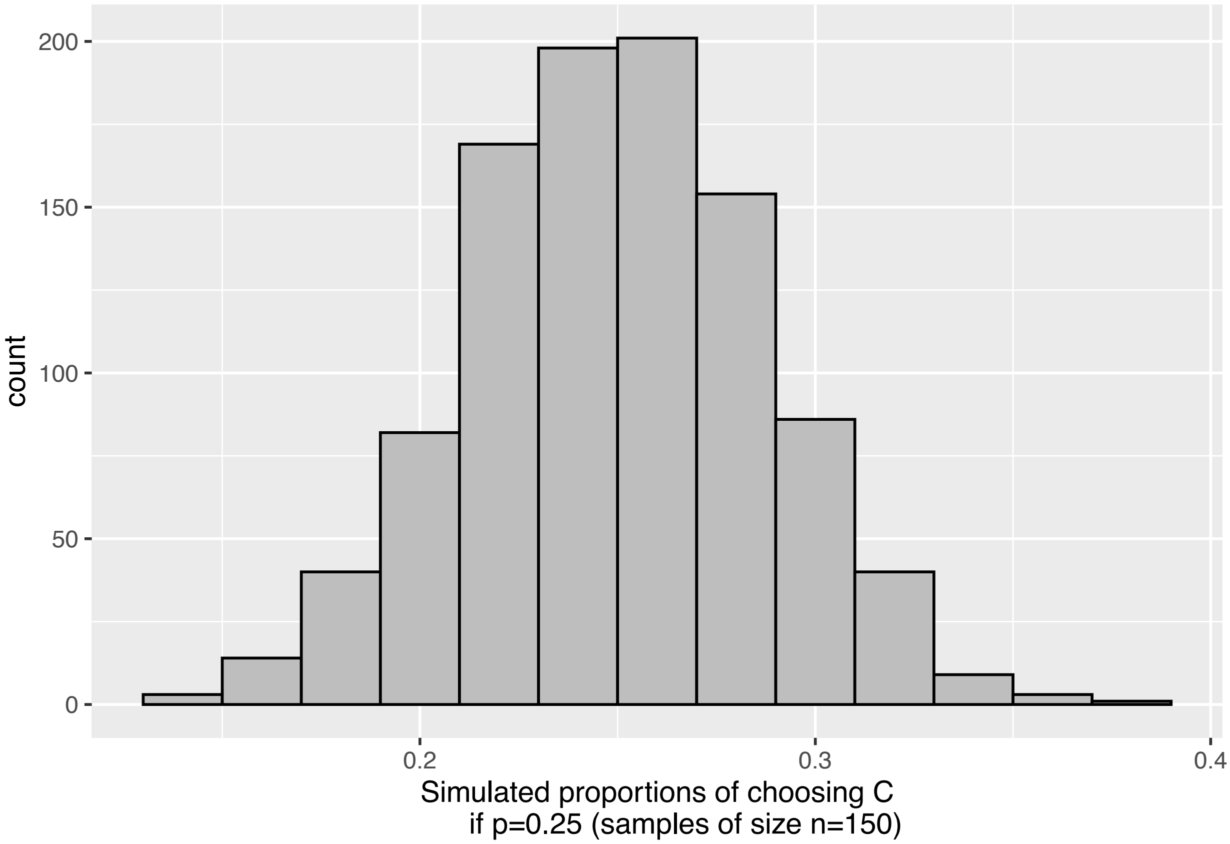
(c) Use R to estimate the sampling distribution of the test statistic under the assumption that people choose multiple choice answers randomly when guessing. Use 1000 repetitions and set the seed to the last 3 digits of your student number. Generate the plot of this estimated sampling distribution and describe the distribution in a few sentences. How does this compare to your sketch from part (b) of this question?

```
n_observations <- 150
repetitions <- 1000
simulated_stats <- rep(NA, repetitions)
set.seed(136)

#generate simulation
for (i in 1:repetitions){
choose_sim <- sample(c("A", "B", "C", "D"),
                    size = n_observations,
                    prob = c(0.25,0.25,0.25,0.25),
                    replace = TRUE)
  choose_p <- sum(choose_sim == "C") / n_observations
  simulated_stats[i] <- choose_p;
}

#save simulation
choose <- data_frame(p_C = simulated_stats)

#plot simulation
choose %>% ggplot(aes(x = p_C)) +
  geom_histogram(binwidth = 0.02, colour = "black", fill = "grey") +
  xlab("Simulated proportions of choosing C
       if p=0.25 (samples of size n=150)")
```



```
choose_pvalue <- choose %>%
  filter(abs(p_C - 0.25) >= abs((34/150) - 0.25)) %>%
  summarise(choose_pvalue = n() / repetitions)
choose_pvalue
```

```
## # A tibble: 1 x 1
##   choose_pvalue
##           <dbl>
## 1         0.549
```

The distribution is symmetrical and unimodal and centered at approximately 0.25. This is approximately the same as I have sketched.

(d) Use the simulation results from (c) to carry out the test at a 5% significance level and make an appropriate conclusion.

Since the significance level is 0.05 and the p-value(0.549) is greater than 0.05, there is no evidence against the null hypothesis. Therefore, students have no preference when choosing A,B,C,or D (the porportion for choosing C is in fact 0.25).