STA130H1S – Winter 2024

Week 4 Practice Problems - Sample Answers

N. Moon and J. Speagle and Christopher Li

Instructions

How do I hand in my solutions and how do I check my work

You will submit your solutions (.Rmd and .pdf) on MarkUs at the following link: https://markus4.teach.cs.toronto.edu/2024-01/courses/1 Submissions are due at 5pm on Thursdays; see Quercus page for the specific deadline for each problem set.

Usually when you do an assignment, you don't find out whether your answers are correct until *after* the deadline, when you get your grade back. However, using MarkUs, you can submit your work before the deadline and run tests to check your solutions!

Note: Some parts of some questions may not be covered by tests in MarkUs, but you're still responsible for reviewing the posted solutions and make sure you understand them. Some of these parts will also be graded in some weeks.

Some of the optional questions have tests in MarkUs, but you won't be penalized for not completing these (or failing the tests for these parts) when we grade your work after the submission deadline. The tests for these parts are provided for your guidance.

What to do if a test fails on MarkUs

- Take a deep breath! Your work won't really be graded until the deadline, so start early to make sure you have lots of time to resolve issues before the deadline.
- Read the message to get hints about what the problem is. For example "variable X not present" means that you may have a typo in the name of the variable we're looking for - re-read the question carefully and make sure you're following the instructions.
- Search on Piazza to see if other classmates have encountered a similar error (and if not, consider posting a screenshot of the error message).
- Come to TA or instructor office hours with your issue.

IMPORTANT - SETUP

```
# Instructions for how students should define tests
STUDENT_NUMBER <- 1010057028; # Replace this number by your real student number
```

Question 1: Canadian Legal System

A criminal court considers two opposing claims about a defendant: they are either **innocent** or **guilty**. In the Canadian legal system, the role of the prosecutor is to present convincing evidence that the defendant is not innocent. Lawyers for the defendant attempt to argue that the evidence is *not convincing enough* to rule out that the defendant could be innocent. If there is not enough evidence to convict the defendant and they are set free, the judge generally does not deliver a verdict of "innocent", but rather of *"not guilty"*.

Please answer each of the questions below in 1-3 sentences using your own words.

(a) If we look at the criminal trial example in the hypothesis test framework, what would be the **null hypothesis** and what would be the **alternative**?

> The null hypothesis would be "guilty" in this situation, and the alternative hypothesis would be "not guilty." This is because "not guilty" is a rejection of "guilty," so it's a rejection of the null hypothesis.

(b) In the context of this problem, describe what **rejecting the null hypothesis** would mean.

> Rejecting the null hypothesis (the defendent is not guilty) means that the judge has enough evidence to declare that the defendent is not guilty.

(c) In the context of this problem, describe what **failing to reject the null hypothesis** would mean.

> Failing to reject the null hypothesis (the defnedent is guilty) means that there wasn't enough evidence to declare that the defendent is not guilty.

(d) In the context of this problem, describe what a **Type II error** would be.

> A Type 2 Error happens when we fail to reject the null hypothesis but we should have instead. In this context, it means we wrongfully convicted someone who should've been let free.

(e) In the context of this problem, describe what a **Type I error** would be.

> A Type 1 Error happens when we reject the null hypothesis but we shouldn't have. In this contxt, it means we have let someone go free when they should've been convicted guilty.

(f) In general, the Canadian legal system is designed to preferentially *avoid* what types of errors?

- A: Type I errors more often than Type II errors
- B: Type II errors more often than Type I errors
- C: Type I and Type II errors at equal rates
- D: Neither Type I nor Type II errors (it's something else)

```
# Replace NULL below by your answer ("A", "B", "C", or "D")
Q1f_choice <- "B"
```

Describe what your answer above means in your own words.

> We should focus on reducing the amount of wrongful convictions but we should also try and reduce the amount of people who walk free when they shouldn't.

(g) [OPTIONAL] In a criminal case, the standard of proof needed for a guilty verdict is "proof beyond a reasonable doubt". What α-level do you think this corresponds to and why? (As a reminder: an α-level is the threshold used to reject the null hypothesis in favour of the alternative when the p-value $p < \alpha$. It controls the Type I error rate.)

The alpha level will probably be around 0.01 (extremely certain) to 0.05 (standard used for statistical signifigance).

(h) [OPTIONAL] In a civil case, the standard of proof needed for someone to be found liable is "on balance of probabilities". What α-level do you think this corresponds to and why?

The alpha level will probably be around 0.05 (standard used for statistical signifigance) to 0.10 (semi-confident that it is signifigant).

Question 2: Instagram Usage

Roughly 38% of the world's 5.3 billion active internet users regularly access Instagram. Consider the following scenario:

- Suppose that the Department of Statistical Sciences (DoSS) is conducting a study to see if this percentage/fraction (38%, or $f_{\text{pop}} = 0.38$) is the **same** among their undergraduate students (i.e. all students in an undergraduate statistics program that is run by DoSS) or potentially **different** (either higher or lower).
- Suppose $n = 600$ DoSS students are **randomly selected** and asked whether or not they use Instagram regularly ("yes" or "no").
- Suppose that $n_{\text{yes}} = 252$ of these $n = 600$ students respond that they use Instagram. This now gives an observed test statistic of $\hat{f}_{\text{stu}} = n_{\text{yes}}/n$.

(a) What is the **null hypothesis** H_0 in terms of the **test statistic**, the fraction f_{stu} of students who regularly use Instagram? What is the **alternative hypothesis** H_1 in terms of H_0?

Please write your answer below using math notation including H_0, H_1, f_{stu}, and/or f_{pop}.

$H_0 : f_{\text{stu}} = f_{\text{pop}}$, $H_1: f_{\text{stu}} \neq f_{\text{pop}}$

Now, please write a sentence describing the claims of the null and alternative hypotheses without any math notation (H_0, H_1, f_{stu}, and/or f_{pop}).

> The null hypothesis is that the frequencey of students who use instagram is the same as the frequency of people (in general) who use instagram. The alternative hypothesis is that the frequency of students who use instagram is not the same as the frequency of people (in general) who use instagram.

(c) Below is R code that simulates the number of students who use Instagram in **ONE random sample** of $n = 600$ DoSS students **under the null hypothesis**.

Note 1: ***You must include the `set.seed(STUDENT_NUMBER)` line in your code and run it every time you run the cell below to ensure the sample is fully reproducible.*** *This controls the "random number seed" that R uses to generate a list of pseudo-random numbers, which guarantees you will get the same set of "random" numbers each time. Normally the random number seed is set automatically when you initialize R by your computer using semi-random numbers such as the current time in milliseconds. For additional information, check out this video.*

Note 2: Because each of you will have a different student number, you will all have slightly different answers.

```
### DO NOT CHANGE THE CODE BELOW ####
print(STUDENT_NUMBER)      # Make sure that this is your student number
```

```
## [1] 1010057028
```

```
                           # If it isn't correct, go to the SETUP section and
                           # store the correct value of your student number in the
                           # STUDENT_NUMBER variable (and run the code chunk)
                           # then come back to this question
                           # Make sure you don't share screenshots of your
                           # solution that share your student number with others
set.seed(STUDENT_NUMBER)   # REQUIRED so the random sample is reproducible!

# setup
n_sample <- 600  # number of observations in random sample
options <- c("Yes", "No")  # options to respond
p_1 <- 0.38  # probability of option 1
```

```
p_options <- c(p_1, 1 - p_1)  # probabilities for both options

# sample!
responses <- sample(options, size=n_sample, prob=p_options, replace=TRUE)
### DO NOT CHANGE THE CODE ABOVE ####
```

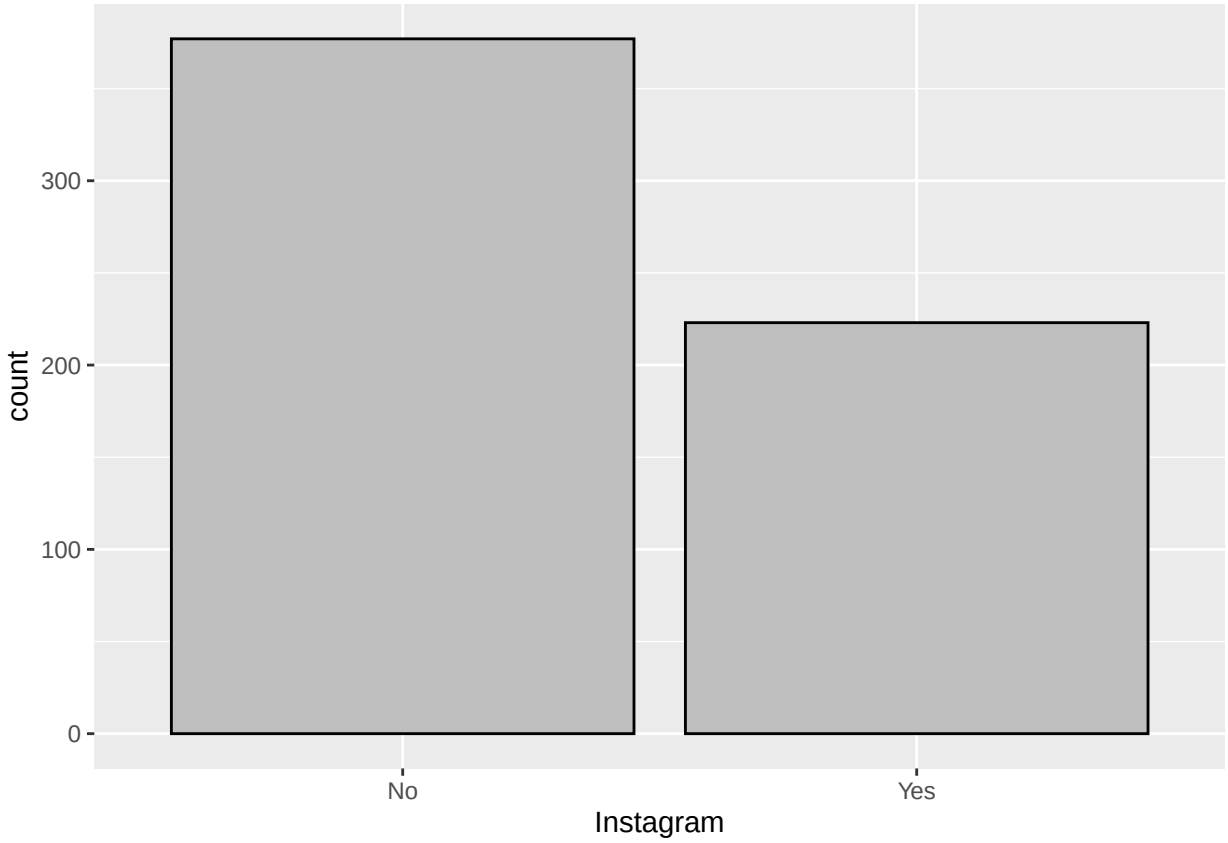
How many regular Instagram users do you have in your simulated sample of students?

```
# Replace NULL below by your answer
Q2c_number_of_instagram_users_in_sample_of_DoSS <- 600
```

(d) [OPTIONAL] Use `geom_bar()` to visualize the number of students who regularly use Instagram versus those who don't from your simulated sample with a bar plot.

Hint: You can make a vector a column of a `tibble` *like this:* `tibble(flips = c("Head", "Tail", "Tail"))`.

```
data.frame(resp = responses) %>% ggplot(aes(x=resp)) +
  geom_bar(color = "black",
           fill = "gray") +
  labs(x = "Instagram")
```



(e) Use the simulated `responses` vector to compute the **simulated test statistic** $\hat{f}_{\text{stu,sim}}$, the *simulated* fraction of DoSS students who use Twitter.

```
# Replace NULL below by your answer
Q2e_fhat_stu_sim <- sum(responses == "Yes") / n_sample
```

How does the simulated fraction $\hat{f}_{\text{stu,sim}}$ compare to the assumed student fraction under the null hypothesis

f_{stu} and to the observed fraction $\hat{f}_{\text{stu}}$ from the $n = 600$ sampled DoSS students?

```
# Replace NULL below by your answer
Q2e_f_stu <- 0.38
Q2e_fhat_stu <- 0.42
```

Please describe any differences you observe between the three numbers above in 1-3 sentences below.

> Our simulation gave us a 37% usage rate in students, and in reality there's an about 38% usage rate in the general public. If we soley look at students however, we get a 42% usage rage among students.

(f) The code below simulates the **sampling distribution** of the test statistic $\hat{f}_{\text{stu}}$, i.e. the distribution of the simulated values $\hat{f}_{\text{stu,sim}}$ under the null hypothesis H_0.

Note: Make sure that you saved your student number in ***STUDENT_NUMBER*** *in the setup section at the top of this document and ran that code by clicking on the green arrow in the top right corner of the code chunk, before doing this question.*

```
### DO NOT CHANGE THE CODE BELOW ####
print(STUDENT_NUMBER)      # Make sure that this is your student number
```

```
## [1] 1010057028
```

```
                           # If it isn't correct, go to the SETUP section and
                           # store the correct value of your student number in the
                           # STUDENT_NUMBER variable (and run the code chunk)
                           # then come back to this question
                           # Make sure you don't share screenshots of your
                           # solution that share your student number with classmates
set.seed(STUDENT_NUMBER + 2)  # Do not change this line

# setup
n_trials <- 1000  # number of trials
n_sample <- 600  # number of observations in random sample
options <- c("Yes", "No")  # options to respond
p_options <- c(0.38, 1 - 0.38)  # probabilities for both options

# simulate!
fhat_stu_simulations <- rep(NA, n_trials)
for (i in 1:n_trials){
  responses <- sample(options, size=n_sample, prob=p_options, replace=TRUE)
  fhat_stu_sim <- sum(responses == "Yes") / n_sample
  fhat_stu_simulations[i] <- fhat_stu_sim
}
### DO NOT CHANGE THE CODE ABOVE ####
```

What are the simulated values that comprise the sampling distribution?

Hint: The answer you are looking for is already calculated in the simulation above.

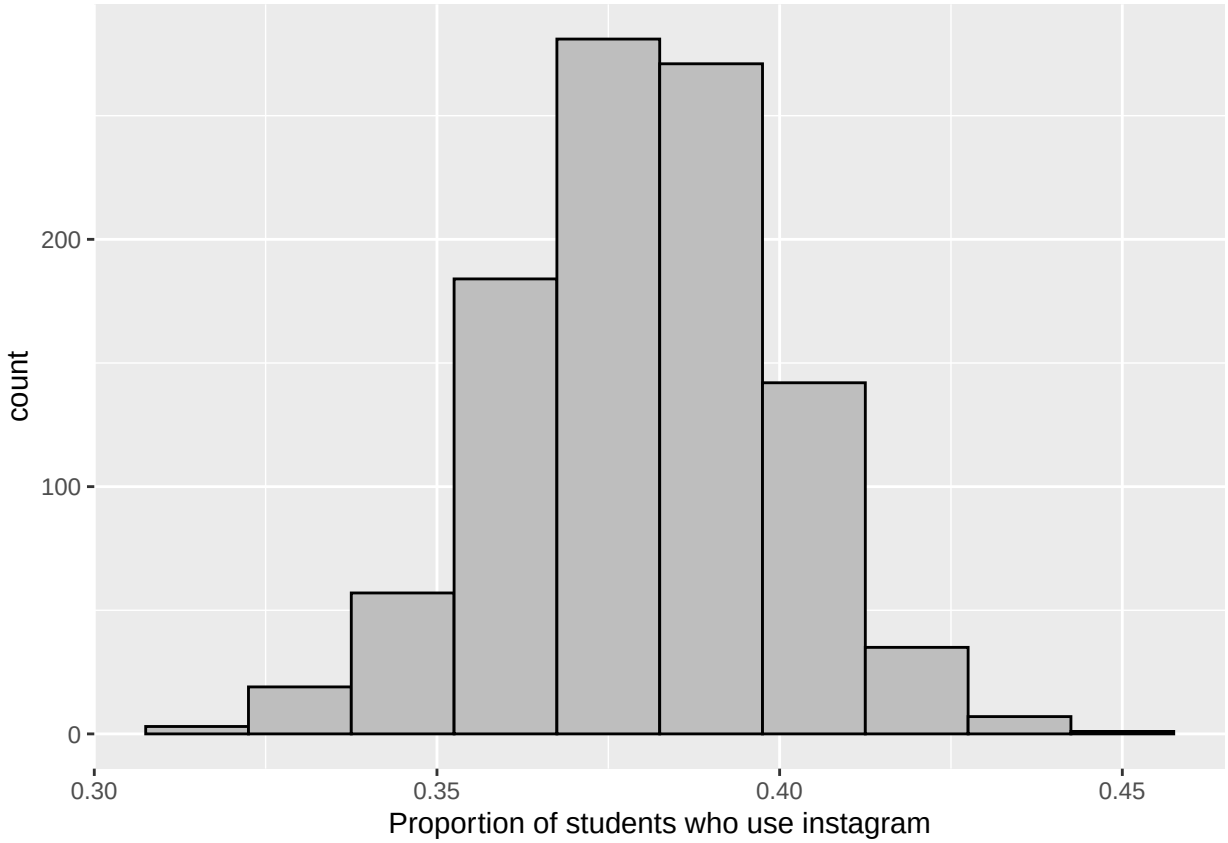
```
# Replace NULL with your answer below
Q2f_vector_of_simulated_test_statistics <- fhat_stu_simulations
```

What are its mean and standard deviation?

```
# Replace NULL with your answer below
Q2f_mean_of_simulated_test_statistics <- mean(fhat_stu_simulations)
Q2f_stddev_of_simulated_test_statistics <- sd(fhat_stu_simulations)
```

(g) [**OPTIONAL**] Make an appropriate plot to visualize your simulated sampling distribution.

```
data.frame(data = fhat_stu_simulations) %>% ggplot(aes(x = data)) +
  geom_histogram(bins=10, color="black", fill="gray") +
  labs(x = "Proportion of students who use instagram")
```



Describe the distribution in 1-2 sentences using some of the key words/phrases we have learned over the past few weeks.

> This distribution is unimodal and approximately normal. It has a center at roughly 0.375 and a range of approximate 0.15. There does not seem to be any tails.

(h) What is the definition of a ***p*-value**?

> Probability that achieving the observed statistic (or worse) is from chance alone.

Please comment on the difference between a **1-sided** versus a **2-sided** p-value and explain why the p-value associated with the null hypothesis H_0 and alternative H_1 hypothesis in this question is 2-sided rather than 1-sided.

> 1 sided p values only look at 1 side of the distribution (why it's called one sided). The alternative hypothesis is usually $>$ or $<$. 2 sided p values look at both sides of the distriubtion, so the alternative hypothesis is usally !=.

(i) Compute the p-value based on the null hypothesis H_0 using the values of $\hat{f}_{\text{stu,sim}}$ you computed from the sampling distribution, the observed test statistic $\hat{f}_{\text{stu}}$, and the assumed true value under the null hypothesis f_{stu}.

Hint: Remember you can take advantage of "coercion" in R to compute the number of objects that satisfy a logical condition and `abs()` *to compute the absolute value of a value/vector.*

```
# Replace NULL with your answer below
Q2i_pvalue_2sided <- sum(abs(fhat_stu_simulations - 0.38) >= abs(0.42 - 0.38)) / n_trials
```

(j) Consider a **rejection rule** where we will reject the null hypothesis H_0 if our p-value falls below a pre-specified α level (i.e. $p < \alpha$). At the $\alpha = 0.05$ significance level, should we reject the null in favour of the alternative based on the computed p-value?

```
# Replace NULL with your answer below (either TRUE or FALSE)
alpha_level <- 0.05
Q2j_reject <- TRUE
```

(k) Please select the correct statement regarding the interpretation of the p-value computed earlier and write 1-2 sentences justifying your answer. "My computed p-value is..."

- A: "...the probability that the proportion of DoSS students who regularly use Instagram matches the general population."
- B: "...the probability that the proportion of DoSS students who regularly use Instagram does not match the general population."
- C: "...the probability of obtaining a number of students who regularly use Instagram in a sample of 600 students at least as extreme as the result in this study."
- D: "...the probability of obtaining a number of students who regularly use Instagram in a sample of 600 students at least as extreme as the result in this study, if the prevalence of regular Instagram users among all DoSS students matches the general population."

```
# Replace NULL by your answer ("A", "B", "C", or "D")
Q2k_choice <- "D"
```

(l) [**OPTIONAL**] Using the slightly-modified seed value in `set.seed()` below, re-compute your p-value and redo your hypothesis test.

```
### DO NOT CHANGE THE CODE BELOW ####
print(STUDENT_NUMBER)      # Make sure that this is your student number
```

```
## [1] 1010057028
```

```
                           # If it isn't correct, go to the SETUP section and
                           # store the correct value of your student number in the
                           # STUDENT_NUMBER variable (and run the code chunk)
                           # then come back to this question
                           # Make sure you don't share screenshots of your
                           # solution that share your student number with classmates
set.seed(STUDENT_NUMBER + 3)  # Do not change this line

# setup
n_trials <- 1000  # number of trials
n_sample <- 600  # number of observations in random sample
options <- c("Yes", "No")  # options to respond
p_options <- c(0.38, 1 - 0.38)  # probabilities for both options

# simulate!
fhat_stu_simulations <- rep(NA, n_trials)
for (i in 1:n_trials){
```

```
  responses <- sample(options, size=n_sample, prob=p_options, replace=TRUE)
  fhat_stu_sim <- sum(responses == "Yes") / n_sample
  fhat_stu_simulations[i] <- fhat_stu_sim
}
### DO NOT CHANGE THE CODE ABOVE ####
```

```
# Replace NULL with your answer below
Q2l_pvalue_2sided <- sum(abs(fhat_stu_simulations - 0.38) >= abs(0.42 - 0.38)) / n_trials

# Replace NULL with your answer below
alpha_level <- 0.05
Q2l_reject <- FALSE
```

How much do the results change?

My results did not change at all.

(m) [**OPTIONAL**] Using the slightly modifed code below, re-compute your p-value but now using 100 times more trials for both the original random number seed and the slighty-modified seed value.

```
### DO NOT CHANGE THE CODE BELOW ####
print(STUDENT_NUMBER)      # Make sure that this is your student number
```

```
## [1] 1010057028
```

```
                           # If it isn't correct, go to the SETUP section and
                           # store the correct value of your student number in the
                           # STUDENT_NUMBER variable (and run the code chunk)
                           # then come back to this question
                           # Make sure you don't share screenshots of your
                           # solution that share your student number with classmates
set.seed(STUDENT_NUMBER + 2)  # Do not change this line

# setup
n_trials <- 100000  # number of trials
n_sample <- 600  # number of observations in random sample
options <- c("Yes", "No")  # options to respond
p_options <- c(0.38, 1 - 0.38)  # probabilities for both options

# simulate!
fhat_stu_simulations <- rep(NA, n_trials)
for (i in 1:n_trials){
  responses <- sample(options, size=n_sample, prob=p_options, replace=TRUE)
  fhat_stu_sim <- sum(responses == "Yes") / n_sample
  fhat_stu_simulations[i] <- fhat_stu_sim
}
### DO NOT CHANGE THE CODE ABOVE ####
```

```
# Replace NULL with your answer below
Q2m_pvalue_2sided_v1 <-
  sum(abs(fhat_stu_simulations - Q2e_f_stu) >=
      abs(Q2e_fhat_stu - Q2e_f_stu)) / n_trials

# Replace NULL with your answer below
alpha_level <- 0.05
Q2m_reject_v1 <- Q2m_pvalue_2sided_v1 < alpha_level
```

```
Q2m_pvalue_2sided_v1
```

```
## [1] 0.04822
```

```
Q2m_reject_v1
```

```
## [1] TRUE
```

```
### DO NOT CHANGE THE CODE BELOW ####
print(STUDENT_NUMBER)     # Make sure that this is your student number
```

```
## [1] 1010057028
```

```
                          # If it isn't correct, go to the SETUP section and
                          # store the correct value of your student number in the
                          # STUDENT_NUMBER variable (and run the code chunk)
                          # then come back to this question
                          # Make sure you don't share screenshots of your
                          # solution that share your student number with classmates
set.seed(STUDENT_NUMBER + 3)  # Do not change this line

# setup
n_trials <- 100000  # number of trials
n_sample <- 600  # number of observations in random sample
options <- c("Yes", "No")  # options to respond
p_options <- c(0.38, 1 - 0.38)  # probabilities for both options

# simulate!
fhat_stu_simulations <- rep(NA, n_trials)
for (i in 1:n_trials){
  responses <- sample(options, size=n_sample, prob=p_options, replace=TRUE)
  fhat_stu_sim <- sum(responses == "Yes") / n_sample
  fhat_stu_simulations[i] <- fhat_stu_sim
}
### DO NOT CHANGE THE CODE ABOVE ####
```

```
# Replace NULL with your answer below
Q2m_pvalue_2sided_v2 <- sum(abs(fhat_stu_simulations - 0.38) >= abs(0.42 - 0.38)) / n_trials

# Replace NULL with your answer below
alpha_level <- 0.05
Q2m_reject_v2 <- TRUE
```

How much do the results change now?

We rejected the null hypothesis.

[OPTIONAL] Question 3: Scottish Medicine

*Note: This entire question is **optional**.*

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 12 of the 12 t-shirts! You can read more about this experiment here.

(a) Without conducting a simulation, describe in 1-2 sentences 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 randomly.

**Hint: You might be able to draw some inspiration from some of the examples discussed in Week 4's class meeting.*

The sampling distribution will be approximately normal with a center at 6.

(b) Write down a hypothesis test that the woman being a lucky guesser (our null hypothesis, H_0) as opposed to having some ability to correctly identify Parkinson's disease by smell (the alternative hypothesis H_1).

Hint: Think carefully about whether you are performing a 1-sided or 2-sided hypothesis test here. How would you interpret your alternative hypothesis in each case?

Null hypothesis is that the woman truly has a 50% chance of guessing correctly. Alternative hypothesis is that the woman has >50% chance of guessing correctly.

(c) Carry out a simulation study and plot the simulated sampling distribution under the null hypothesis H_0. Then, compute the p-value given that the woman correctly classified 12 of 12 t-shirts (i.e. our observed test statistic).

Hint: You should be able to re-use a lot of your code from Question 2 here!

```r
set.seed(STUDENT_NUMBER + 3)  # REQUIRED so the results are reproducible!

# write your code below
# setup
n_trials <- 100000  # number of trials
n_sample <- 12  # number of observations in random sample
options <- c("Yes", "No")  # options to respond
p_options <- c(0.5, 1 - 0.5)  # probabilities for both options

# simulate!
fhat_tshirt_simulations <- rep(NA, n_trials)
for (i in 1:n_trials){
  responses <- sample(options, size=n_sample, prob=p_options, replace=TRUE)
  fhat_tshirt_sim <- sum(responses == "Yes") / n_sample
  fhat_tshirt_simulations[i] <- fhat_tshirt_sim
}

one_sided <- sum(fhat_tshirt_simulations >= 1) / n_trials
```

(d) Given an $p < \alpha = 0.05$ rejection rule, what is your conclusion about this hypothesis test based on your computed p-value?

The woman is most likely able to smell Parkisan's disease.

(e) Actually, initially 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. It was only eight months later that the final individual was diagnosed with the disease. Assuming only 11 of 12 guesses being correct, what is the new associated p-value?

Hint: You should not have to re-run a whole new set of simulations to compute this new p-value. If you do decide to resimulate, please remember to include `set.seed(STUDENT_NUMBER + 3)` *at the beginning of the*

code block.

```
# write your code below
one_sided_2 <- sum(fhat_tshirt_simulations >= 11/12) / n_trials
```

(f) Based on our $p < \alpha = 0.05$ rejection rule, does the conclusion of the hypthesis test the same for an observed test statistic of 11/12 compared an observed test statistic of 12/12?

The conclusion is still the same.