Module 4 synchronous class code (complete)

Video code and additional template code can be found in the 'Big Doc of Sample Code'. Use that link to pull the file in, it will be called STA130-F21-module-4-extras.

In this code demo, we are going to work through a test for one proportion.

1. State your hypotheses

$$H_0 : p_C = 0.25$$
$$H_A : p_C \neq 0.25$$

In plain words: Our null hypothesis is that the chance of picking C is that same as for any of the other options. There were 4 choices, so if students are picking randomly, we'd expect to see C chosen about 1 in 4 times, 0.25 (25%). Our alternative is that the true proportion of times students pick C is NOT 25%.

2. Calculate your test statistic (real world)

2A. Load `tidyverse` and data

```
library(tidyverse)
```

2B. Save an object that says how many observations are in our sample data

```
n_observations <- 196
```

2C. Calculate the test statistic

```
test_stat <- 71/196 #some calculation using my real world data
test_stat
```

```
## [1] 0.3622449
```

3. Simulate under the null hypothesis

3A. Set values for simulation

```r
# Suppose the last three digits of my student ID were 123 and I was asked
# to use the last three digits for my set seed
set.seed(123)
repetitions <- 100
simulated_stats <- rep(NA, repetitions)
```

3B. Automate simulation with a `for` loop (simulation world)

```r
for (i in 1:repetitions){
  new_sim <- sample(c("C", "not C"), size = 196, replace = TRUE, prob = c(0.25, 0.75))

  sim_val <- sum(new_sim == "C")/n_observations

  # store this stat in its own little slot in the storage vector we made in 3C.
  simulated_stats[i] <- sim_val
}
```

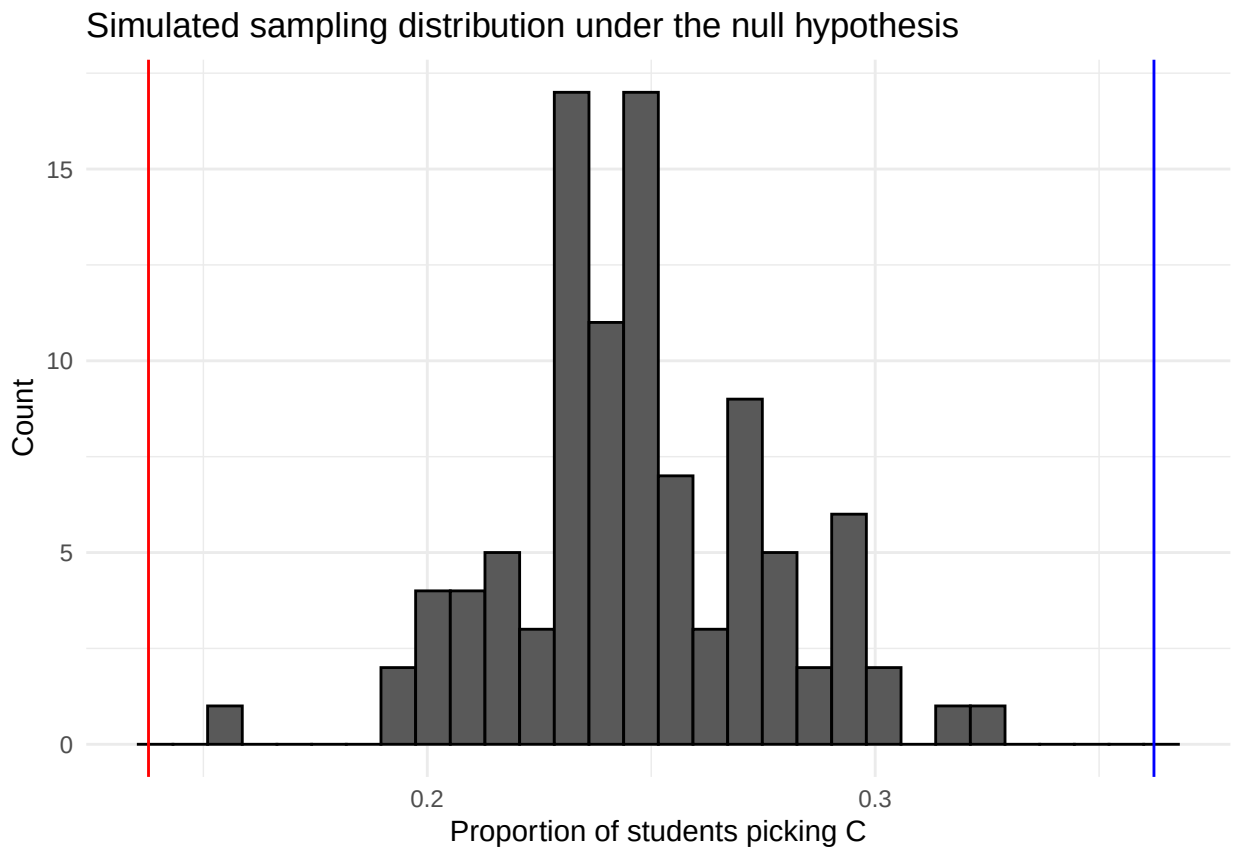
3C. Turn results into a data frame so we can use ggplot for plotting

```r
sim_tibble <- tibble(simulated_statistics = simulated_stats)
```

4. Evaluate the evidence against the null hypothesis

```r
# the value from your null hypothesis
hypothesized_value <- 0.25

ggplot(sim_tibble, aes(x = simulated_statistics)) +
  geom_histogram(bins = 30, colour = "black") +
  labs(x="Proportion of students picking C", y="Count",
       title = "Simulated sampling distribution under the null hypothesis") +
  geom_vline(xintercept = hypothesized_value - abs(test_stat-hypothesized_value),
             colour = "red") +
  geom_vline(xintercept = hypothesized_value + abs(test_stat-hypothesized_value),
             colour = "blue") +
  theme_minimal()
```




```
p_value <- sim_tibble %>%
  filter(simulated_statistics <= hypothesized_value - abs(test_stat-hypothesized_value) |
         simulated_statistics >= hypothesized_value + abs(test_stat-hypothesized_value)) %>%
  summarise(p_value = n() / repetitions) %>%
  as.numeric()
p_value
```

```
## [1] 0
```

Our p-value is 0.

5. Make a conclusion

Strength of evidence conclusion

We have very strong evidence against the hypothesis that students are randomly picking any option when presented with a question they have not ability to answer with any knowledge (at least in the case where there are 4 options, no stakes, the question is written in Windings and the students are the type that take STA130 and turn up to class).

What type of error are we at risk of making if we use a significance level of 0.1?

At the significance level of $\alpha = 0.1$, we would reject this hypothesis and so would be at risk of making a type I error, that is, it could be the case that it is in fact true overall that students pick randomly, and we just observed a particularly unusual test statistic by chance, causing us to incorrectly reject the null.

Some guidelines for how small is small? This table tells you how to comment on the **strength of evidence against** H_0.

P-value	Evidence
p-value > 0.10	no evidence against H_0
$0.05 <$ p-value < 0.10	weak evidence against H_0
$0.01 <$ p-value < 0.05	moderate evidence against H_0
$0.001 <$ p-value < 0.01	strong evidence against H_0
p-value < 0.001	very strong evidence against H_0