Lab 9

Taylor Blair

Math 141, Week 10

Due: Before your Week 11 lab meeting

Goals of this lab

1. Practice another hypothesis testing example.
2. Practice computing probability distributions.
3. Practice finding conditional probabilities.
4. Practice working with random variables.

Problems

```r
# Insert libraries here
library(tidyverse)
library(infer)
library(reticulate)
matplotlib <- import("matplotlib")
matplotlib$use("Agg", force = TRUE)
library(ggplot2)
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import math
```

Problem 1

Let X = household size for owner-occupied housing units in the US. Based on the US Census, the following table contains the probability function for X.

x	1	2	3	4	5	6	7
p(x)	0.217	0.363	0.165	0.145	0.067	0.026	0.017

a. The following category is actually "7 or more". Explain why we have to cut it off at 7 to make X a random variable. What impact does cutting X off at 7 have the mean and standard deviation of the random variable?

- *Why is it cut off at 7*

  - Privacy reasons
  - Skew
  - Looking at the trends above, any value beyond 8 is likely to be less than one percent.
  - "Or more" can't really know
- *What happens to mean and SD?*
  - Although housing with 7 or more makes up only 1.7%, the larger a family the more the curve is shifted right.
  - Standard deviation will also increase as the formula for SD is distance from the mean, and the change will create outiers.

---

b. Is X a discrete or continuous random variable? Justify your answer.

- **Discrete**
  - There are a finite number of variables
  - Last I checked there weren't households with half a person, so it can't be continuous.

---

c. If we sum up the probabilities of each outcome of X, what value will we get?

- 1, as it represents a percentage proportion.

---

d. What is the probability that a unit has five or less people in it?

- $1 - (0.217 + 0.363 + 0.165 + 0.145 + 0.067) = 0.043 = 4.3\%$

---

e. What is the mean household size for owner-occupied houses in the US?

```
households <-c(1:7)
weights_house <- c(0.217, 0.363, 0.165, 0.145, 0.067, 0.026, 0.017)
weighted.mean(households, weights_house)
```

```
## [1] 2.628
```

- Probably not part of a person in a household. So for realism sake, best to round to 3.
- The or more implies that it will be further skewed right, but there is no way to know by how much without more data.

Problem 2

A national survey conducted by the Pew Research Center in late 2018 produced the following estimates about educational attainment and Twitter use among US adults:

- 10% have less than a high school diploma; 8% of these adults use Twitter
- 59% have a high school diploma but no college degree; 20% of these adults use Twitter
- 31% have a college degree; 30% of these adults use Twitter

a. Suppose we have a random sample of 1000 US adults. Based on the information above, distribution these adults into the following table:

Table	Twitter User	None Twitter User	Total
Less than high school	8	92	100
High school	118	472	590
College	93	217	310
Total	219	781	1000

---

b. What percentage of US adults who use Twitter have less than a high school diploma?

- $\frac{8}{219} \approx 3.652\%$

---

c. What percentage of US adults who use Twitter have a high school degree but no college degree?

$\frac{118}{219} \approx 53.881\%$

---

d. What percentage of US adults who use Twitter have a college degree?

- $\frac{93}{219} \approx 42.465\%$

---

e. When we condition on being a Twitter user, which educational groups are more likely than they were when we didn't condition on anything? Which are less likely?

- The most common in the real world were individuals with a only a high school diploma (59%).
- On twitter the most common group was individuals with only high school diplomas (53.881%)
- Uncondtioned the most common group were college individuals as they have a 30% usage.

Problem 3

Let's return to the exam handing back problem from the probability worksheet but suppose Donna collected their exam and so we only have three exams to hand back and we again hand them back randomly. Let X = number of students who get their correct exam.

a. Find the probability function for X.

```
def single_combination(grab_list):
  combination = []
  grab_from = np.arange(len(grab_list)).tolist()
  for y in range(0,len(grab_list)):
```

```
    combination.append(grab_from[grab_list[y]])
    grab_from.pop(grab_list[y])
  return combination

def update_grab_list(grab_list):
  edit_num = 0
  for z in range (0,len(grab_list)):
    if edit_num==z:
      if(grab_list[z]+1)%(len(grab_list)-z)<(grab_list[z]+1):
          edit_num +=1
      grab_list[z]=(grab_list[z]+1)%(len(grab_list)-z)
  return grab_list

def make_combinations(tests):
  grab_list = np.zeros(tests, dtype=np.int).tolist()
  master_list = []
  for x in range (0,math.factorial(tests)):
    master_list.append(single_combination(grab_list))
    grab_list = update_grab_list(grab_list)
  return master_list

def evalute_combinations(combinations):
  eval_nums = []
  for x in range (0, len(combinations)):
    eval_nums.append(evaluate_scenario(combinations[x]))
  return eval_nums

def stat_graph(correct_tests):
  y = correct_tests
  x = np.arange(len(correct_tests)) # make array here
  plt.bar(x, y)
  plt.xlabel('Correct Placements')
  plt.ylabel('Number of Tests')
  plt.title('Distbution of Tests')
  plt.show()

def evaluate_scenario(test_list):
  tests_correct = 0
  for x in range (0,len(test_list)):
    if test_list[x]==x:
      tests_correct+=1
  return tests_correct

def stat_math(correct_tests):
  percents = np.array(correct_tests, dtype=float)
  percents = percents/np.sum(correct_tests)*100
  stats = pd.DataFrame(columns = ['Correctly Delivered Tests', "Total", "Percent"])
  stats["Correctly Delivered Tests"] = np.arange(len(correct_tests))
  stats["Total"] = correct_tests
  stats["Percent"] = percents
  print(stats)
  print("Mean number correct tests: ", np.average(np.arange(len(correct_tests)),
  weights=correct_tests))
```

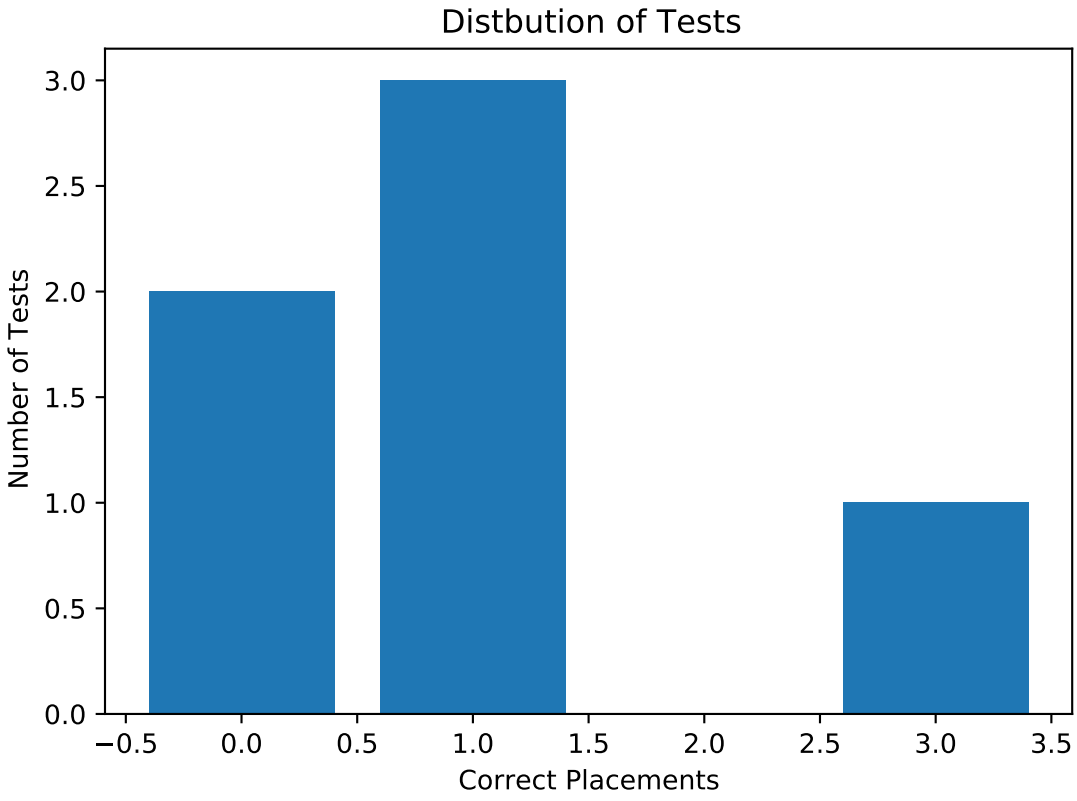
```
def stat_process(num_tests, results):
  processed_stats = []
  for x in range (0,num_tests+1):
    total_num = 0
    for y in range (0, len(results)):
      if results[y]==x:
        total_num+=1
    processed_stats.append(total_num)
  return processed_stats

def total_stats(stat_list, num_tests):
  simple_stats = stat_process(num_tests, stat_list)
  stat_graph(simple_stats)
  stat_math(simple_stats)

def all_scenarios(tests):
  all_combinations = make_combinations(tests)
  eval_nums = evalute_combinations(all_combinations)
  total_stats(eval_nums, tests)


all_scenarios(3)
```

```
##    Correctly Delivered Tests  Total    Percent
## 0                          0      2  33.333333
## 1                          1      3  50.000000
## 2                          2      0   0.000000
## 3                          3      1  16.666667
## Mean number correct tests:  1.0
```



b. What is the probability that everyone gets their correct exam?

- $\frac{1}{3!} = \frac{1}{6} = 16.\bar{6}\%$

c. What is the probability that at least one person gets their correct exam?

- $1 - \frac{2}{3!} = 1 - \frac{2}{6} = \frac{2}{3} = 66.\bar{6}\%$

d. On average, how many students do we expect to receive their correct exam?

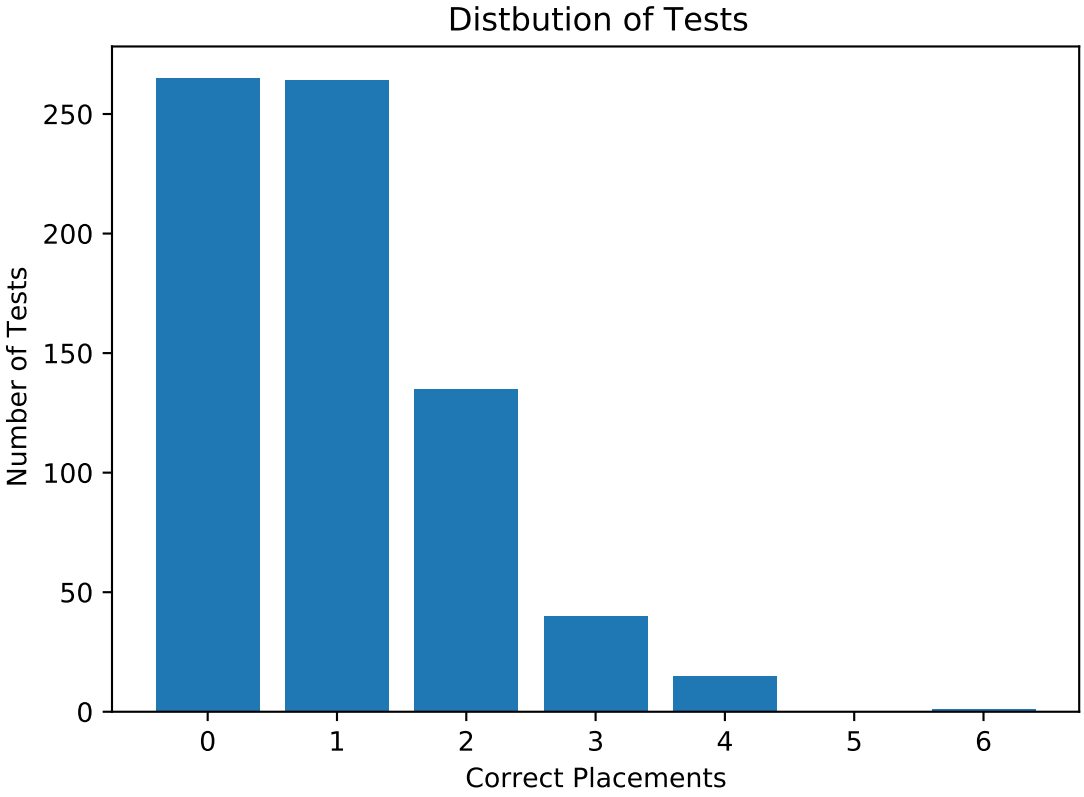
```
correct_tests <-c(0:3)
prop_correct <- c(2, 3, 0, 1)
weighted.mean(correct_tests, prop_correct)
```

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

e. Suppose we actually needed to hand back six exams to six students. Would the probability that all of the students got their own exam back be smaller, the same, or larger than it was for three people? Justify your answer.

```
all_scenarios(6)
```

```
##    Correctly Delivered Tests  Total    Percent
## 0                          0    265  36.805556
## 1                          1    264  36.666667
## 2                          2    135  18.750000
## 3                          3     40   5.555556
## 4                          4     15   2.083333
## 5                          5      0   0.000000
## 6                          6      1   0.138889
## Mean number correct tests:  1.0
```



- $\frac{1}{6!} = \frac{1}{720} = 0.13\bar{8}\%$
- Much much smaller. There is only one correct permutation out of the 6!

Problem 4

As we discussed in class, a common mistake with conditional probabilities is to equate P(A given B) with P(B given A). In this problem, we want you to come up with your own example.

a. Give an example where P(A given B) does not equal P(B given A).

---

1. Professor Kelly at Reed

- The probability of a Professor being named Kelly given they are in the statistics department
  - There are 4 individuals in the stastics department: *Kelly McConville*, Johnathan Wells, Johnathan Kadish, & Tom
  - **The above is based on Kelly verbally confirming**
  - $\frac{1}{4} = 25\%$
- The probability of a professor being in the stats department given they are named Kelly
  - There are two current professors at Reed named Kelly: Kelly N. Chacón (Chemistry) and Kelly McConville (Statistics)
  - Reed Faculty
  - $\frac{1}{2} = 50\%$

2. What are the teens up to these days
   - Based on data from the Youth Risk Survey 2017
   - The probabability of a high schooler respondeding "yes" to the question "*Were Bullied On School Property (during the 12 months before the survey)*" in the youth risk survey given they identify as "gay or lesbian".
     - There were 1246 respondants who identified as gay or lesbian and responded to the question "*Were Bullied On School Property (during the 12 months before the survey)*"
     - Of the individuals who responded, 349 stated they were bullied on campus.
     - $\frac{349}{1246} \approx 28\%$
   - The probabability of a high schooler identifying as "gay or lesbian" given they responded "yes" to the question "*Were Bullied On School Property (during the 12 months before the survey)*" in the youth risk survey
     - There were 14,606 individuals that responded yes to the question "*Were Bullied On School Property (during the 12 months before the survey)*"
     - Of the individuals who responded yes, 349 identify as "gay or lesbian"
     - $\frac{349}{14606} \approx 2.38\%$

---

b. For your example, which probability is more likely? Justify your answer.

---

1. What's in a name
   - The likelihood of being a stats professor at reed given you are named Kelly.
   - There is only one person who is in both a stats professor and named kelly. There are 3 other stats professors opposed to one other Professor Kelly.
   - $\frac{1}{2} > \frac{1}{4}$
2. Risky youth buisness
   - The likelihood of being bullied given you identify as gay
   - The two proportions generated while seemingly similar in theory, are completley different problems. One asks what percent of a population identifies as Gay (typically a very small number), versus what percent of individuals are bullied and gay.
   - $\frac{349}{1246} > \frac{349}{14606}$

---

c. For your example, which probability is *easier* to estimate? Justify your answer.

---

1. Professor Kelly at Reed
   - The data was relatively easy to assertain as there are only 4 individuals in the first probability, and two in the second.
   - Both rational
   - Honestly about the same. I can name all 4 stats professors, but I had to check how many other Kelly's taught at Reed. So maybe being a stats professor given you are named Kelly.
2. I was going to look at sunscreen but they removed that stat so here's a depressing stastic

- I had to estimate the total number of individuals that responded "gay or lesbian" who answered the question "*Were Bullied On School Property (during the 12 months before the survey)*". So that was the more difficult stastic to calculate.
- "*The probabability of a high schooler identifying as "gay or lesbian" given they responded "yes" to the question "Were Bullied On School Property (during the 12 months before the survey)" in the youth risk survey*"
  - Was the easier question to answer. The total number of indivduals who responded yes to the question was a given.

Problem 5

A recent study asked the question "Does tea provide better support for the immune system than coffee?" Black tea contains L-theanaine, an amino acid that is thought to prime Gamma delta T cells, which are important cells for fending off infections.

To test the conjecture, the study participants were randomly assigned to either drink five or six cups of black tea or coffee every day for two weeks. At the end of the two weeks, blood samples were exposed to an antigen and production of interferon gamma, the immune system response, was measured. A higher production of interferon gamma implies a stronger immune system response.

Let's use their data to test the conjecture that tea drinkers will have a stronger immune system response, on average, than coffee drinkers.

```
# Read in data
immune <- read_csv("/home/courses/math141f18/Data/ImmuneTea.csv")
```

a. State the null and alternative hypotheses in symbols and in words.

- H_0 Null hypothesis
  - There is no difference between tea drinkers and coffee drinkers in terms of their immune system response to antigens
- H_a Alternative hypothesis
  - Tea drinkers have a greater immune response in terms of reaction to an antigens than coffee drinkers.
- α
  - 0.05
  - I am one for the classics.

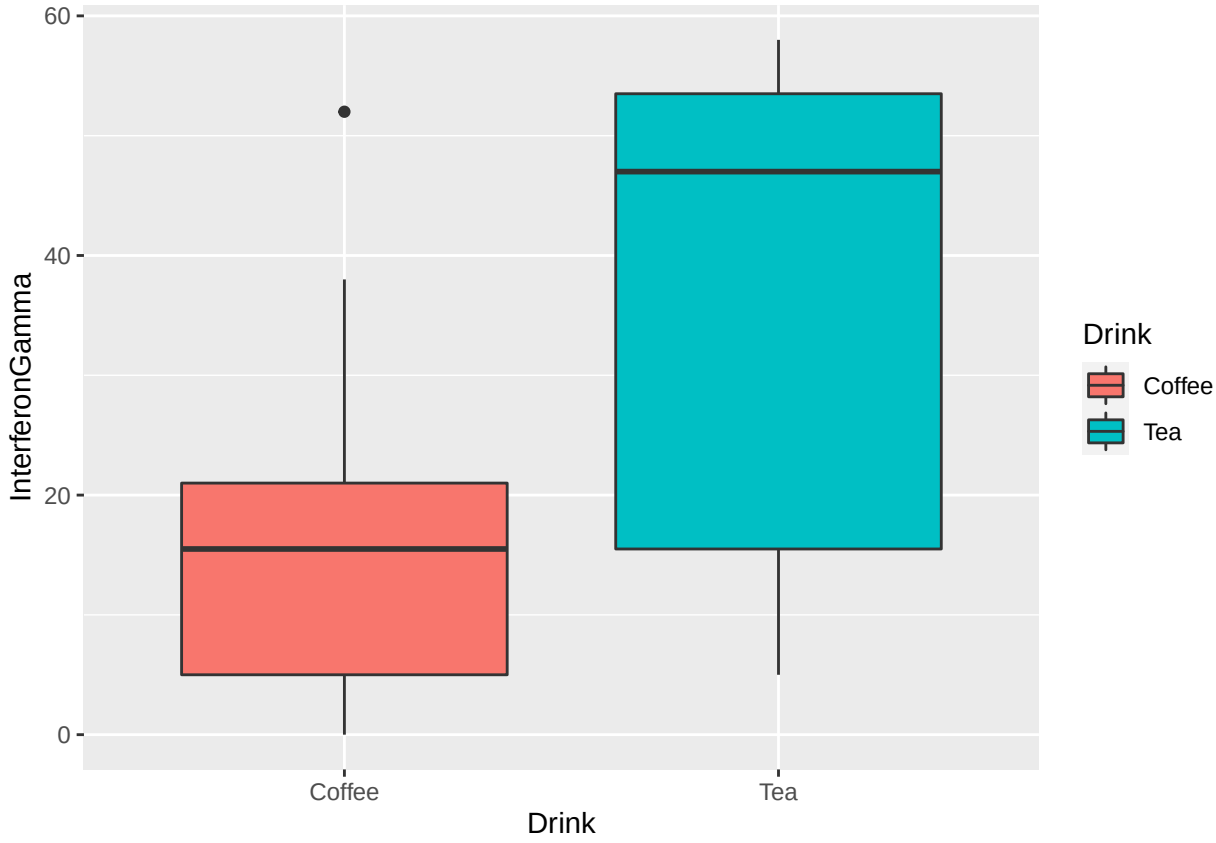
---

b. State what a Type I error and a Type II error represent in the context of the problem.

- Type I
  - False positive
  - Tea drinkers are concluded to have higher immune systems, in reality the null hypothesis was rejected incorrectly.
- Type II
  - False negative
  - The null hypothesis is not rejected, in reality the null ypothesis
- Type III
  - Participants die from drinking 5 to 6 cups of coffee
  - *Joke answer*

c. Construct a useful graph and compute summary statistics. Draw some initial conclusions based on your graph and summary statistics.

```
ggplot(immune, aes(x=Drink, y=InterferonGamma, fill=Drink)) +
  geom_boxplot()
```




```
p_hat_tea <- mean(immune[1][immune[2]=="Tea"])

p_hat_coffee <- mean(immune[1][immune[2]!="Tea"])

test_stat <- p_hat_tea-p_hat_coffee

test_stat
```

```
## [1] 17.11818
```

- Coffee drinkers
  - Have less of an immune system responce, smaller spread, smaller median, skews smaller
- Tea drinkers
  - Greater response, wider spread, skews smaller

---

d. For this problem, explain how we can generate a null distribution through simulation. Make sure to provide clear steps.

- A simulation difference in means can be generated for the null hypothesis.

1. Get two hats, one for tea, one for coffee.
2. Write the slips representing the gamma levels and place them in their respective hats.
3. Take a bootstrap mean for each hat, subtract from one another, repeat for a number of cycles.

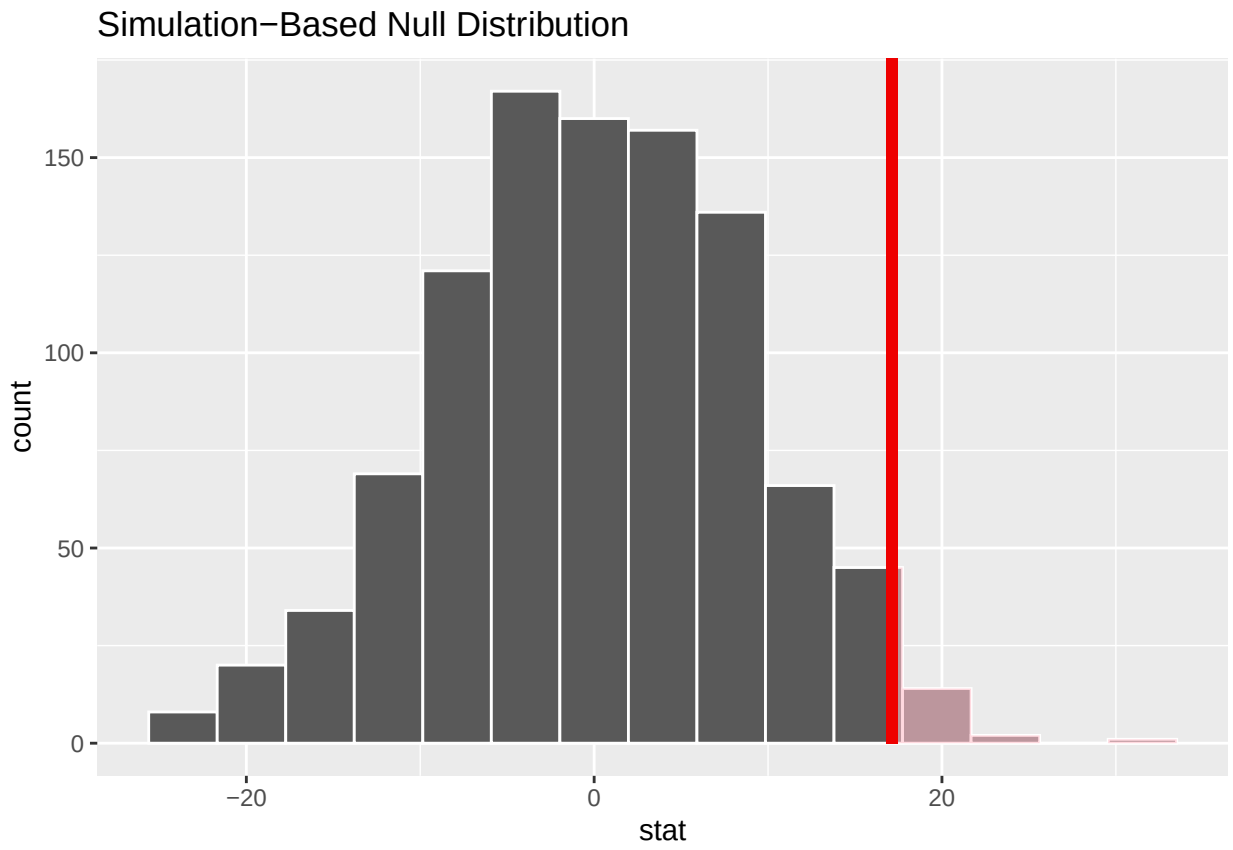
---

e. Generate a null distribution.

```
null <- immune %>%
  specify(explanatory = Drink, response = InterferonGamma) %>%
  hypothesise(null= "independence") %>%
  generate(reps=1000, type="permute") %>%
  calculate("diff in means", order= c("Tea", "Coffee"))
```

f. Compute the p-value.

```
null%>%
  visualise() +
  shade_p_value(obs_stat = test_stat, direction = "right")
```




```
p_immune <- null %>%
  get_p_value(obs_stat = test_stat, direction = "right")

p_immune
```

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

```
## 1    0.02
```

g. Interpret the p-value in the context of the problem.

- The P-value 0.02 is less than 0.05. So It seems likely that I can reject my null hypothesis. In terms of this problem, that means that it is likely there is a difference between immune system response and drink.

---

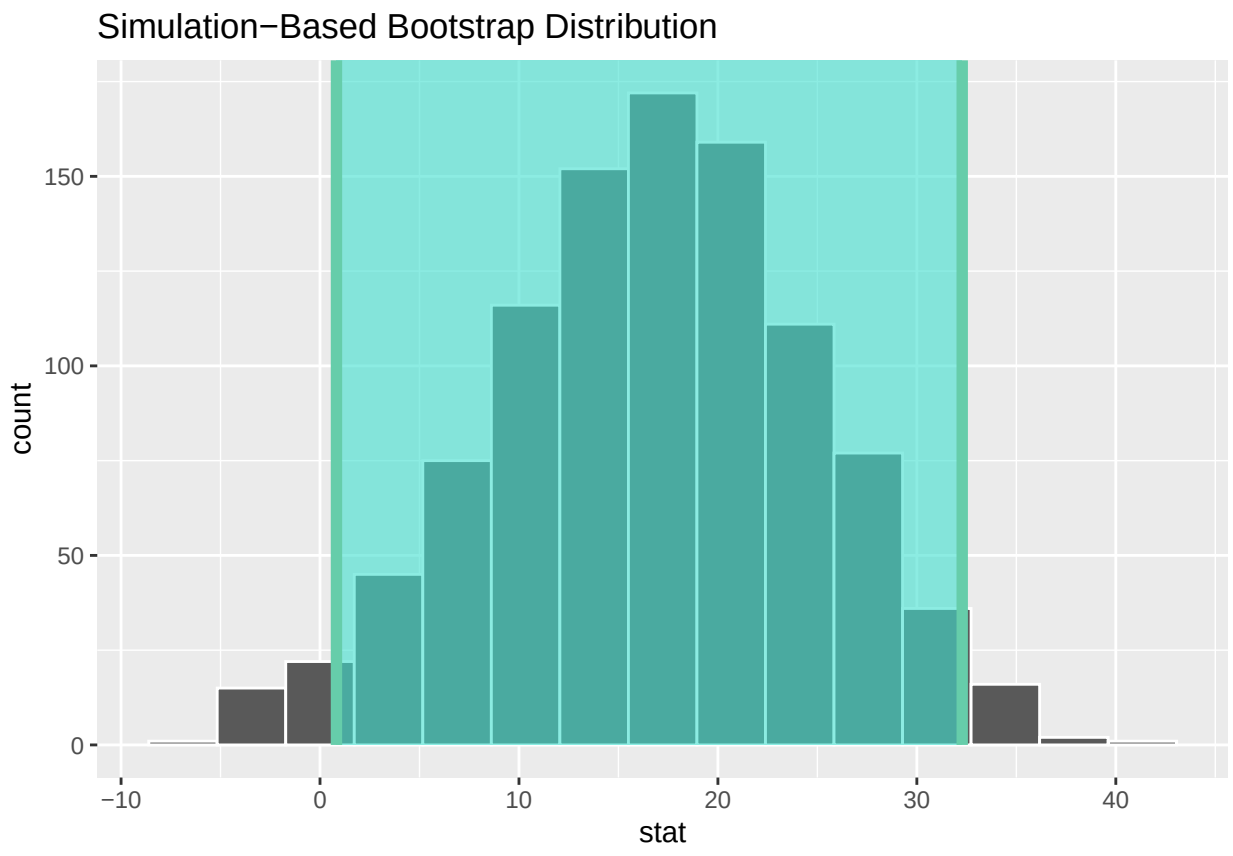
h. Construct a bootstrap distribution for the relevant statistic.

```
boot <- immune %>%
  specify(explanatory = Drink, response = InterferonGamma) %>%
  generate(reps=1000, type="bootstrap") %>%
  calculate("diff in means", order= c("Tea", "Coffee"))
```

i. Compute a 90% confidence interval for the parameter of interest.

```
conf <- boot %>%
  get_confidence_interval()

boot %>%
  visualise() +
  shade_confidence_interval(conf)
```




```
conf
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
```

```
##       <dbl>    <dbl>
## 1    0.827     32.3
```

j. Interpret one of your 90% confidence intervals.

- 90 % confident that drinking tea will raise immune response by (0.827083333333341, 32.2797619047619)

---

k. State some conclusions about the original research question. (Feel free to specify an α level.)

- Given $\alpha = 0.05$ and a p-value of 0.02. I can reject my null hypothesis.
- Of the bootsrapped means, only 2.4% are less than zero. This idicates that I can accept the alternative.
- I would conclude the hypothesis to be true as I can reject the null and accept the hypothesis.
