Homework #04: Bass Mercury

due March 3rd 11:59 PM

Ayush Jain

3/1

Load Packages and Data

```
library(tidyverse)
library(tidymodels)
library(viridis)
```

```
mercury_bass <- read_csv("mercury.csv")
```

Exercise 1

Null Hypothesis: The average mercury level in local bass is 0.46 ppm Alternative Hypothesis: The average mercury level in local bass is greater than 0.46 ppm

$H_0 : mu = 0.46 ppm$ vs. $H_a : mu > 0.46 ppm$

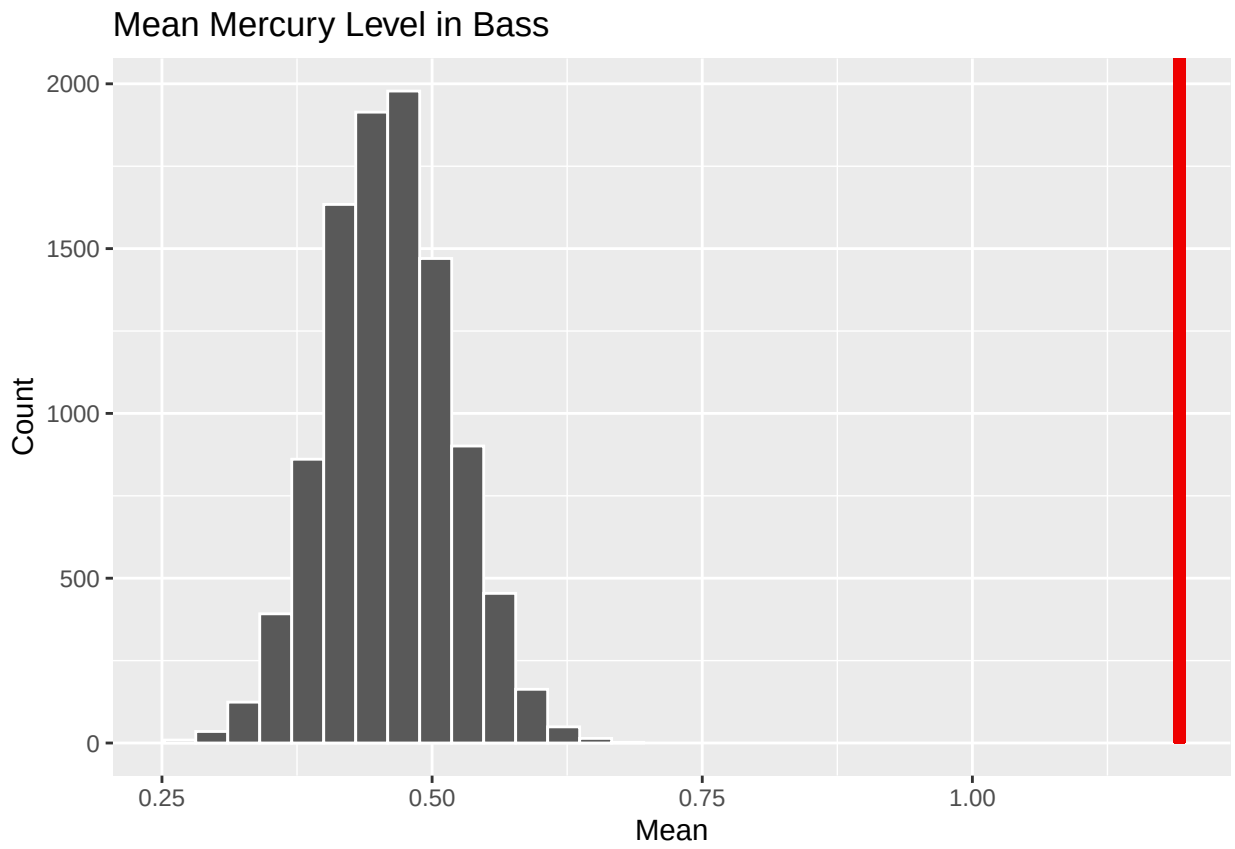
Exercise 2

```
set.seed(2)
 null_dist <- mercury_bass %>%
 specify(response = mercury) %>%
 hypothesize(null = "point", mu = 0.46) %>%
 generate(reps = 10000, type = "bootstrap") %>%
 calculate(stat = "mean")
```

Exercise 3

```
mean_mercury <- mercury_bass %>%
  summarise(mean_mercury = mean(mercury)) %>%
  pull()

visualize(null_dist) +
```

```
  shade_p_value(obs_stat = mean_mercury, direction = "greater") +
  labs(title = "Mean Mercury Level in Bass",
       y = "Count",
       x = "Mean")
```




```
null_dist %>%
  get_p_value(obs_stat = mean_mercury , direction = "greater")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of ‘reps‘ chosen in the ‘generate()‘ step. See
## ‘?get_p_value()‘ for more information.
```

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

Answer: Assuming alpha = 0.05, a p-value of 0 would mean that the test is extremely statistically significant. In this context, this would mean that we can reject the null hypothesis, and further investigate the hypothesis that the average mercury level in local bass is greater than 0.46 ppm

Exercise 4

```
set.seed(4)
mercury_bass %>%
 specify(response = mercury) %>%
 generate(reps = 10000, type = "bootstrap") %>%
 calculate(stat = "mean") %>%
  summarise(lower = quantile(stat, 0.025),
            upper = quantile(stat, 0.975))
```

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1  1.08  1.31
```

Answer: We are 95% confident that the true mean mercury level for local bass is between 1.08 and 1.31 ppm. This is more consistent with the alternative hypothesis, since the alternative hypothesis states that the mercury level is above 0.46 ppm, and 0.46 ppm does not fall within this range(this does not mean that it is necessarily true, it just means that the confidence interval seems to indicate so)

Exercise 5

Null Hypothesis: The average mercury level in bass caught in Waccamaw is the same as bass caught in Lumber Alternative Hypothesis: The average mercury level in bass caught in Waccamaw is not the same as bass caught in Lumber

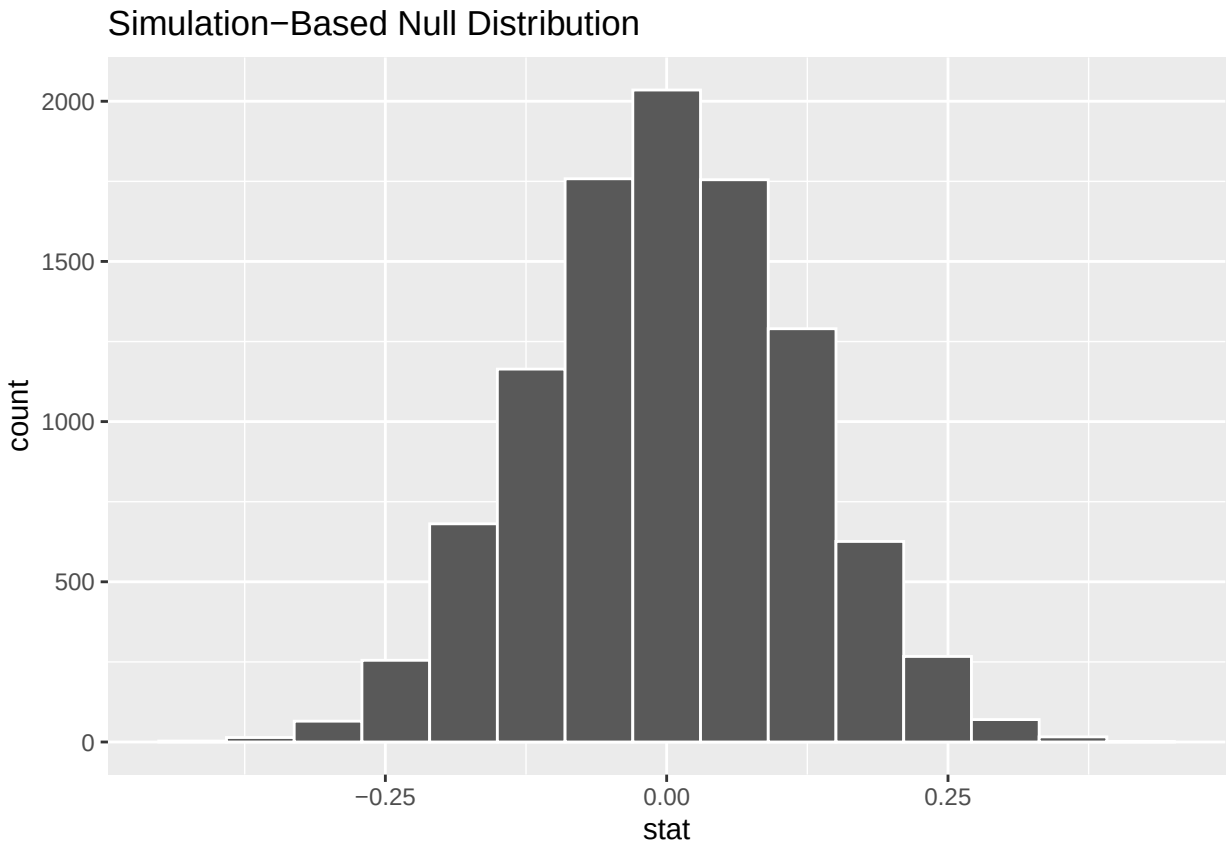
$H_0 : mu_L - mu_W = 0$ vs. $H_a : mu_L - mu_W \neq 0$

```
mercury_bass <- mercury_bass %>%
  mutate(riverName = ifelse (river == 0,
                             "Lumber", "Waccamaw"))
```

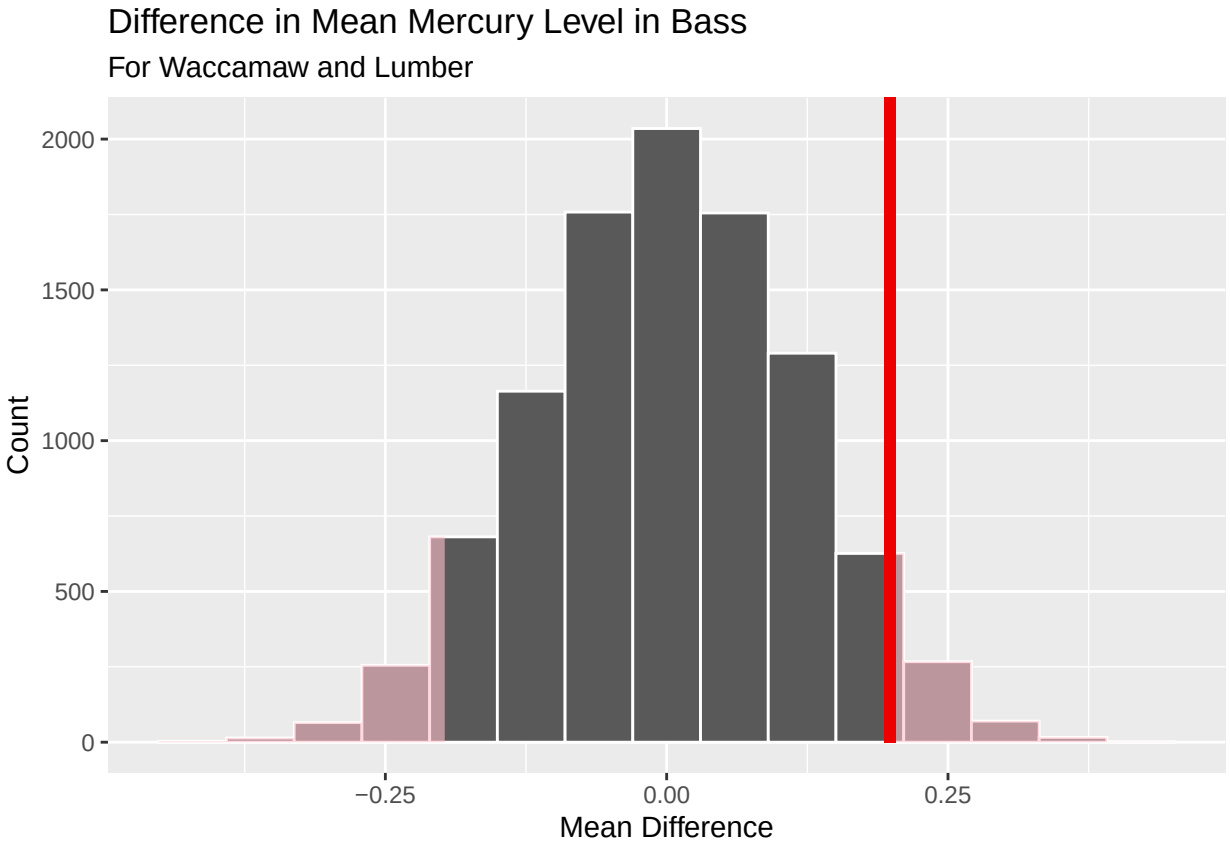
Exercise 6

```
set.seed(6)

null_dist2 <- mercury_bass %>%
 specify(response = mercury, explanatory = riverName) %>%
 hypothesize(null = "independence") %>%
 generate(reps = 10000, type = "permute") %>%
 calculate(stat = "diff in means",
           order = c("Waccamaw", "Lumber"))

mean_mercury_lumber <- mercury_bass %>%
  filter(riverName == "Lumber") %>%
  summarise(mean_mercury_lumber = mean(mercury)) %>%
  pull()
mean_mercury_wacca <- mercury_bass %>%
```

```
  filter(riverName == "Waccamaw") %>%
  summarise(mean_mercury_wacca = mean(mercury)) %>%
  pull()
diff_mean = mean_mercury_wacca - mean_mercury_lumber

visualize(null_dist2)
```




```
visualize(null_dist2) +
  shade_p_value(obs_stat = diff_mean, direction = "two-sided") +
  labs(title = "Difference in Mean Mercury Level in Bass",
       subtitle = "For Waccamaw and Lumber",
       y = "Count",
       x = "Mean Difference")
```




```
null_dist2 %>%
  get_p_value(obs_stat = diff_mean , direction = "two-sided")
```

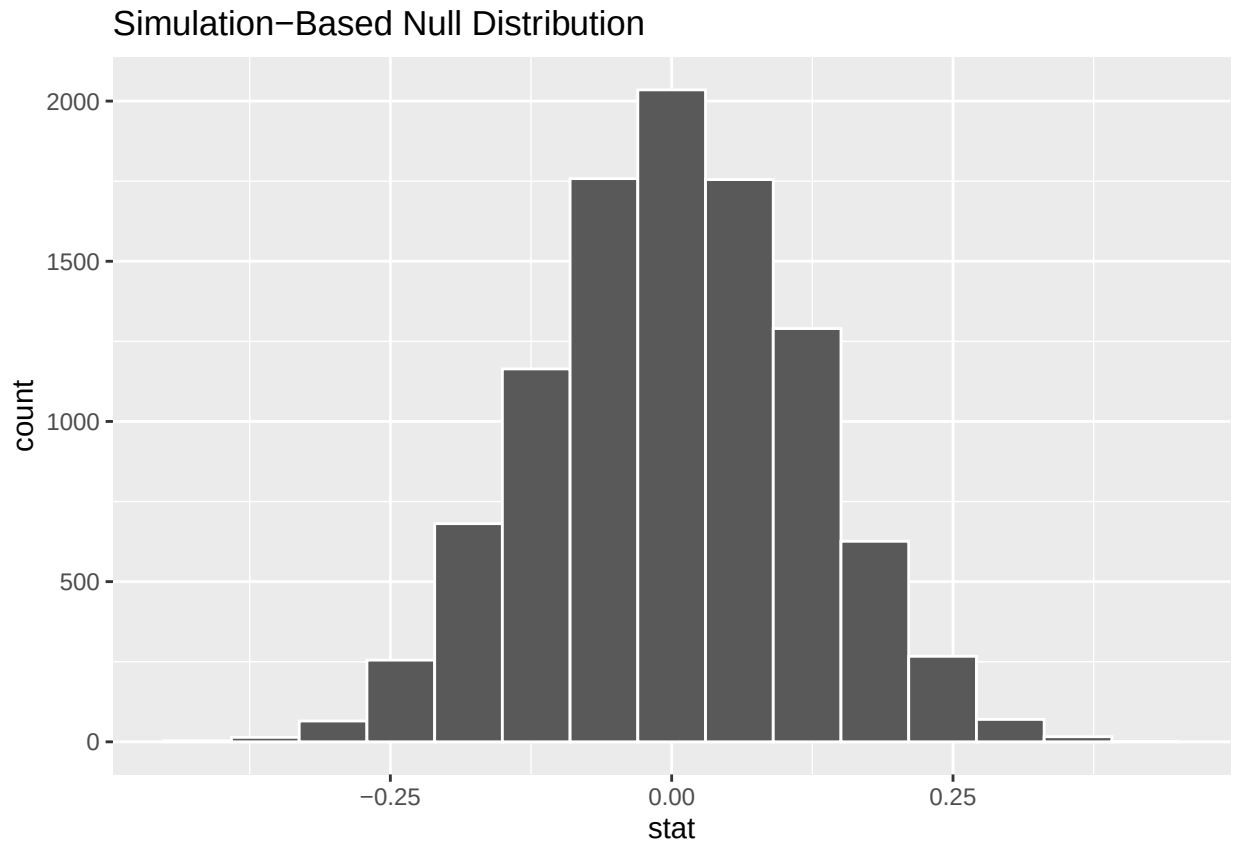
```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0886
```

Answer: The results are not statistically significant since the p-value is greater than 0.05, therefore we fail to reject the null hypothesis i.e. there is no difference in mean mercury levels for bass caught in Waccamaw and Lumber

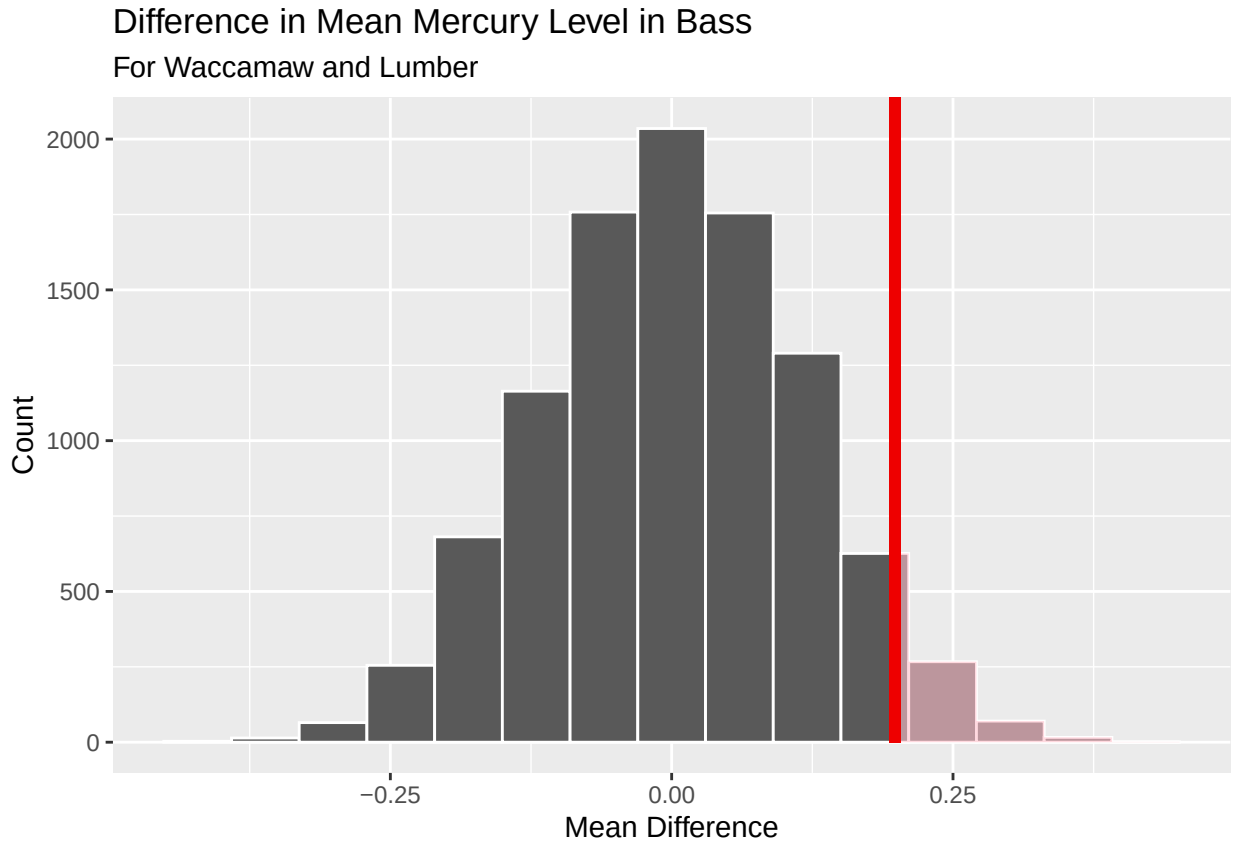
Exercise 7

```
set.seed(6)

null_dist2 <- mercury_bass %>%
 specify(response = mercury, explanatory = riverName) %>%
 hypothesize(null = "independence") %>%
 generate(reps = 10000, type = "permute") %>%
 calculate(stat = "diff in means",
           order = c("Waccamaw", "Lumber"))
```

```
mean_mercury_lumber <- mercury_bass %>%
  filter(riverName == "Lumber") %>%
  summarise(mean_mercury_lumber = mean(mercury)) %>%
  pull()
mean_mercury_wacca <- mercury_bass %>%
  filter(riverName == "Waccamaw") %>%
  summarise(mean_mercury_wacca = mean(mercury)) %>%
  pull()
diff_mean = mean_mercury_wacca - mean_mercury_lumber

visualize(null_dist2)
```




```
visualize(null_dist2) +
  shade_p_value(obs_stat = diff_mean, direction = "greater") +
  labs(title = "Difference in Mean Mercury Level in Bass",
       subtitle = "For Waccamaw and Lumber",
       y = "Count",
       x = "Mean Difference")
```




```
null_dist2 %>%
  get_p_value(obs_stat = diff_mean , direction = "greater")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0443
```

Answer: It is important to specify the null and alternative hypothesis because although we might fail to reject the null hypothesis in lieu of a certain alternative, it is entirely possible that when presented with another hypothesis, we reject the null hypothesis.

As an example, the code above shows what would happen if our alternative hypothesis was that the average mercury content in bass from Waccamaw is more than Lumber. In this case, the p-value (0.0443) is lower than 0.05 and thus statistically significant

Exercise 8

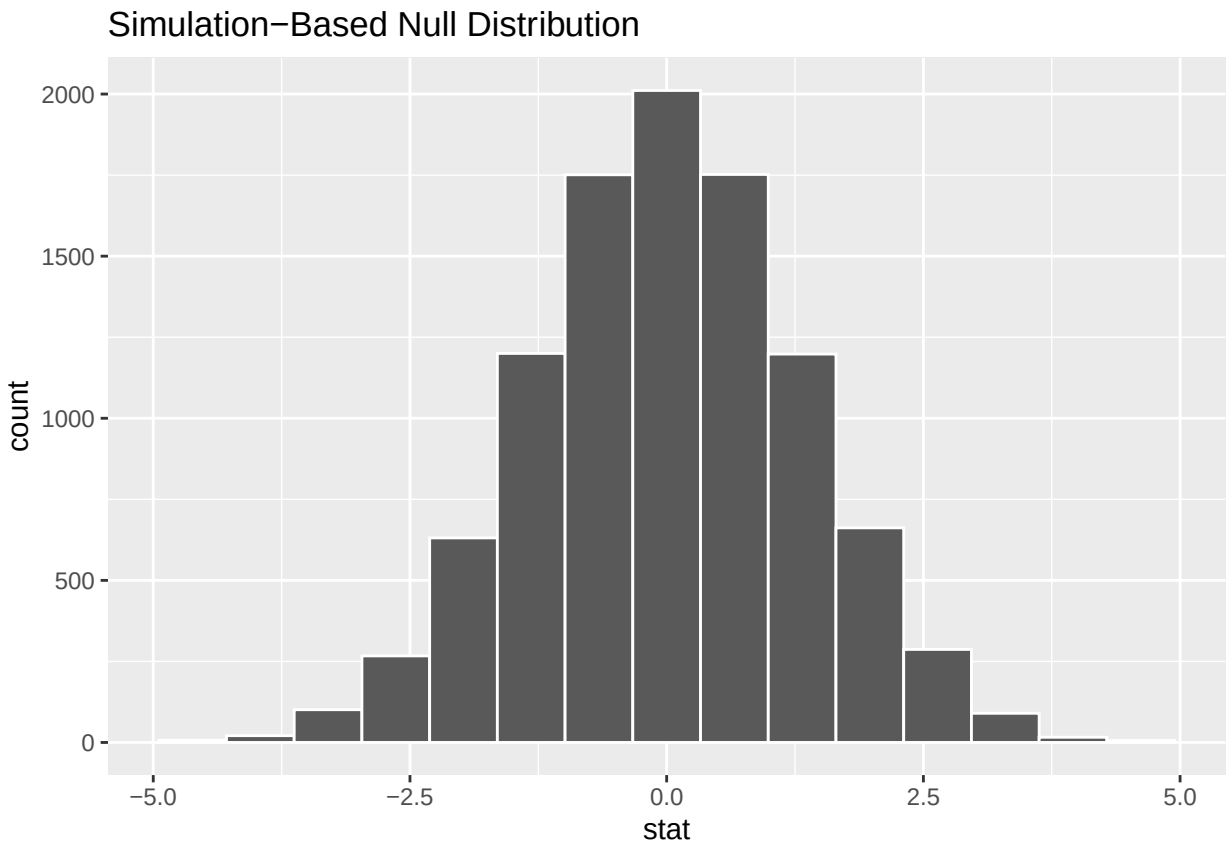
Question: Does the average length differ significantly between bass caught in Waccamaw and bass caught in Lumber river?

Let mu_W be the mean length of Waccamaw bass and mu_L be the mean length of Lumber bass.

Null Hypothesis: The average mercury level in bass caught in Waccamaw is the same as bass caught in Lumber Alternative Hypothesis: The average mercury level in bass caught in Waccamaw is not the same as bass caught in Lumber

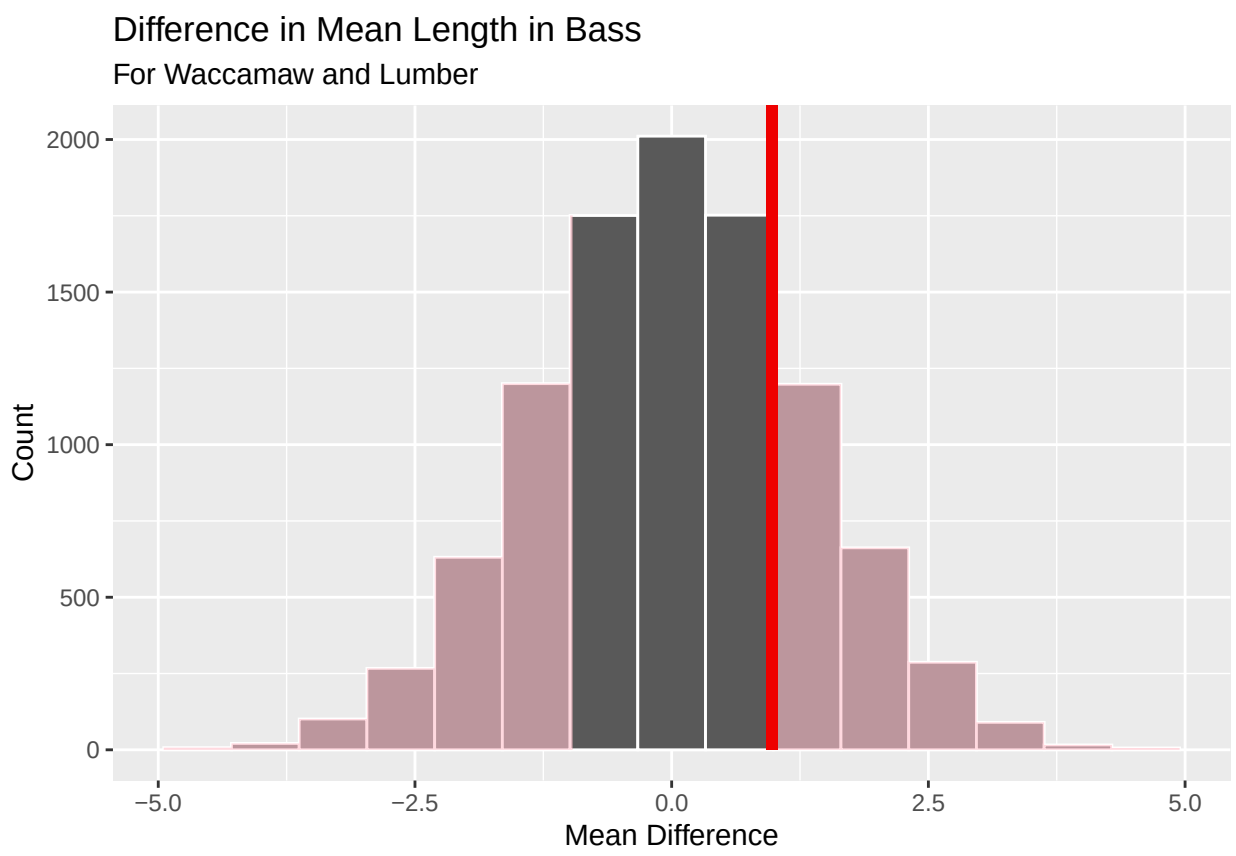
$H_0 : mu_L - mu_W = 0$ vs. $H_a : mu_L - mu_W \neq 0$

```
set.seed(6)

null_dist3 <- mercury_bass %>%
 specify(response = length, explanatory = riverName) %>%
 hypothesize(null = "independence") %>%
 generate(reps = 10000, type = "permute") %>%
 calculate(stat = "diff in means",
           order = c("Waccamaw", "Lumber"))

mean_length_lumber <- mercury_bass %>%
  filter(riverName == "Lumber") %>%
  summarise(mean_length_lumber = mean(length)) %>%
  pull()
mean_length_wacca <- mercury_bass %>%
  filter(riverName == "Waccamaw") %>%
  summarise(mean_length_wacca = mean(length)) %>%
  pull()
diff_meanlength = mean_length_wacca - mean_length_lumber

visualize(null_dist3)
```




```
visualize(null_dist3) +
  shade_p_value(obs_stat = diff_meanlength,
```

```
                direction = "two-sided") +
  labs(title = "Difference in Mean Length in Bass",
       subtitle = "For Waccamaw and Lumber",
       y = "Count",
       x = "Mean Difference")
```




```
null_dist3 %>%
  get_p_value(obs_stat = diff_meanlength,
              direction = "two-sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.459
```

Answer: The p-value indicates that the result is statistically significant, therefore we can reject the null hypothesis that the mean length of fish does not differ between rivers