Homework #04: Bass Mercury

due March 3rd 11:59 PM

Dav King

2/28/2022

Load Packages and Data

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

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

Exercise 1

$H_0 : \mu \leq 0.46$

$H_A : \mu > 0.46$

H_0: The average mercury level is 0.46 ppm or less

H_A: The average mercury level is more than 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

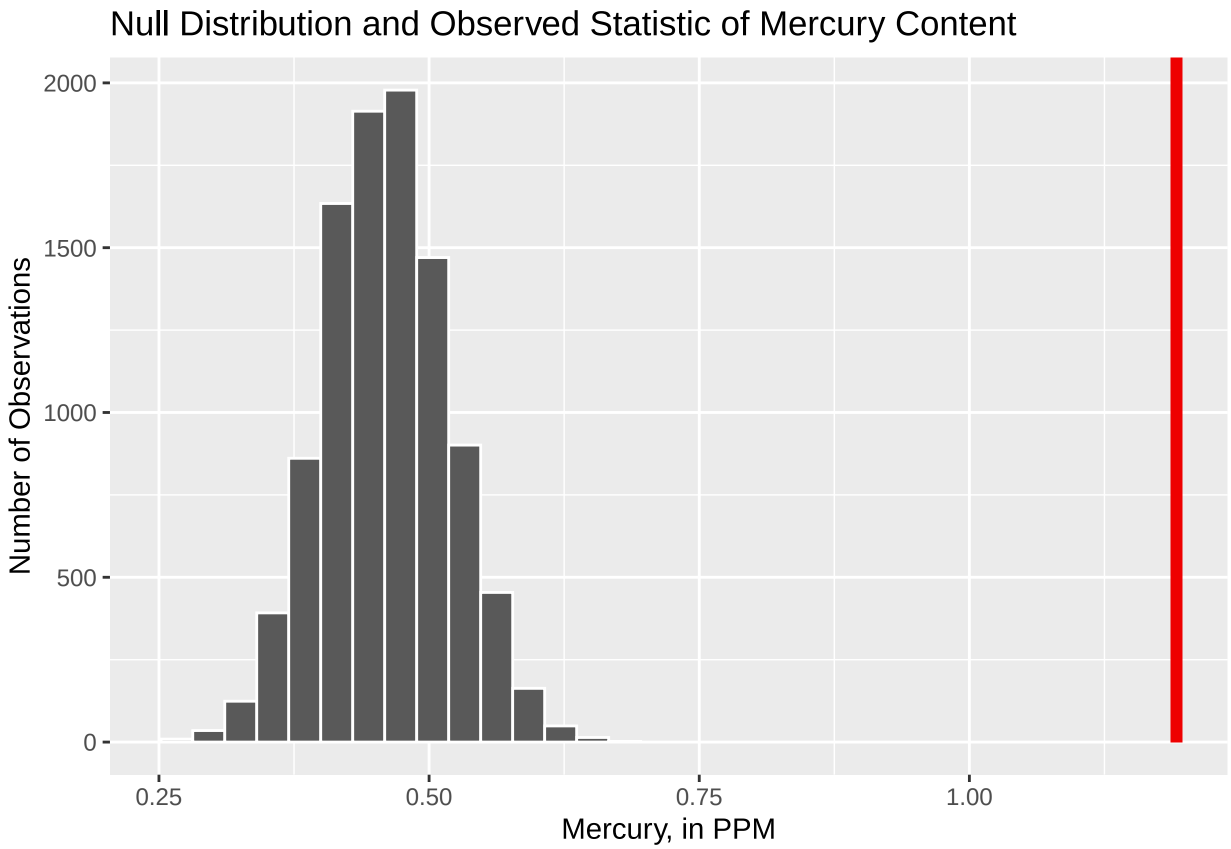
```
pop_mean <- mercury_bass %>%
  summarize(pop_mean = mean(mercury)) %>%
  pull()

null_dist %>%
  get_p_value(obs_stat = pop_mean, 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
```

```
visualize(null_dist) +
  shade_p_value(obs_stat = pop_mean, direction = "greater") +
  labs(title = "Null Distribution and Observed Statistic of Mercury Content",
       x = "Mercury, in PPM", y = "Number of Observations")
```



Assuming $\alpha = 0.05$, no null hypothesis has ever been rejected this hard. Even I didn't get rejected this hard in middle school. We have significant evidence to say that the true population mean is greater than 0.46 ppm mercury in the bass from these two rivers.

Exercise 4

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

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

Given this sample, we are 95% confident that the true population mean mercury content in these two rivers is between 1.08 and 1.31 ppm. This is much more consistent with our alternative hypothesis because, just as it hypothesized, it puts mu at above 0.46 ppm (and H_0 is not even contained on this interval, the other way to calculate our p-value).

Exercise 5

$H_0 : \mu W = \mu L$

$H_A : \mu W \neq \mu L$

H_0: the average mercury content does not differ between bass caught in the two rivers

H_A: the average mercury content differs significantly between bass caught in the two rivers

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

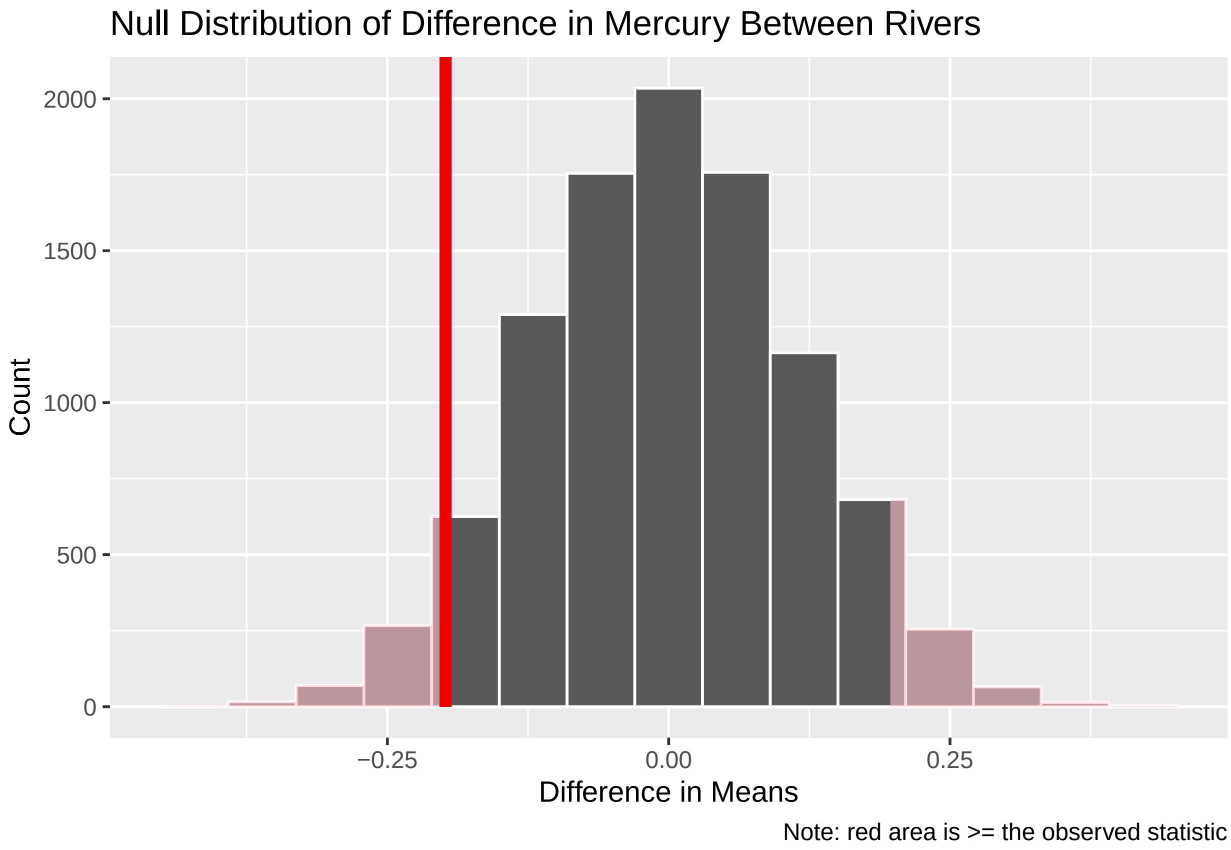
Exercise 6

```
set.seed(6)
null_dist_2 <- mercury_bass %>%
  specify(response = mercury, explanatory = riverName) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means")
```

```
## Warning: The statistic is based on a difference or ratio; by default, for
## difference-based statistics, the explanatory variable is subtracted in the order
## "Lumber" - "Waccamaw", or divided in the order "Lumber" / "Waccamaw" for ratio-
## based statistics. To specify this order yourself, supply ‘order = c("Lumber",
## "Waccamaw")‘ to the calculate() function.
```

```
dif_mean <- mercury_bass %>%
  group_by(riverName) %>%
  summarize(merc = mean(mercury)) %>%
  summarize(dif_mean = merc[riverName == "Lumber"] -
              merc[riverName == "Waccamaw"]) %>%
  pull()

visualize(null_dist_2) +
  shade_p_value(obs_stat = dif_mean, direction = "both") +
  labs(title = "Null Distribution of Difference in Mercury Between Rivers",
       x = "Difference in Means", y = "Count",
       caption = "Note: red area is >= the observed statistic")
```



```
null_dist_2 %>%
  get_p_value(obs_stat = dif_mean, direction = "both")
```

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

At $\alpha = 0.05$, we fail to reject H_0. We do not have significant evidence to suggest that there are different mean levels of mercury in the fish caught in each of the two rivers.

Exercise 7

So we don't HARK and find invalid results that look shiny. We need a hypothesis that we didn't curate to become an explanation for our data - because that makes it entirely free of chance.

```
null_dist_2 %>%
  get_p_value(obs_stat = dif_mean, direction = "less")
```

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

Take this, for example. In our original statistical test, we calculated a p-value of 0.089 - significant at $\alpha = 0.1$, but not at $\alpha = 0.05$. Additionally, we did not have any reason to believe that one river was notably higher than the other - hence, the two-tailed p-test. However, if we (after looking at the data) decided to conduct a one-tailed p-test off of a directional H_A, we would find a p-value significant at $\alpha = 0.05$ - but not necessarily a legitimate one, as we had no reason to investigate that H_A.

Exercise 8

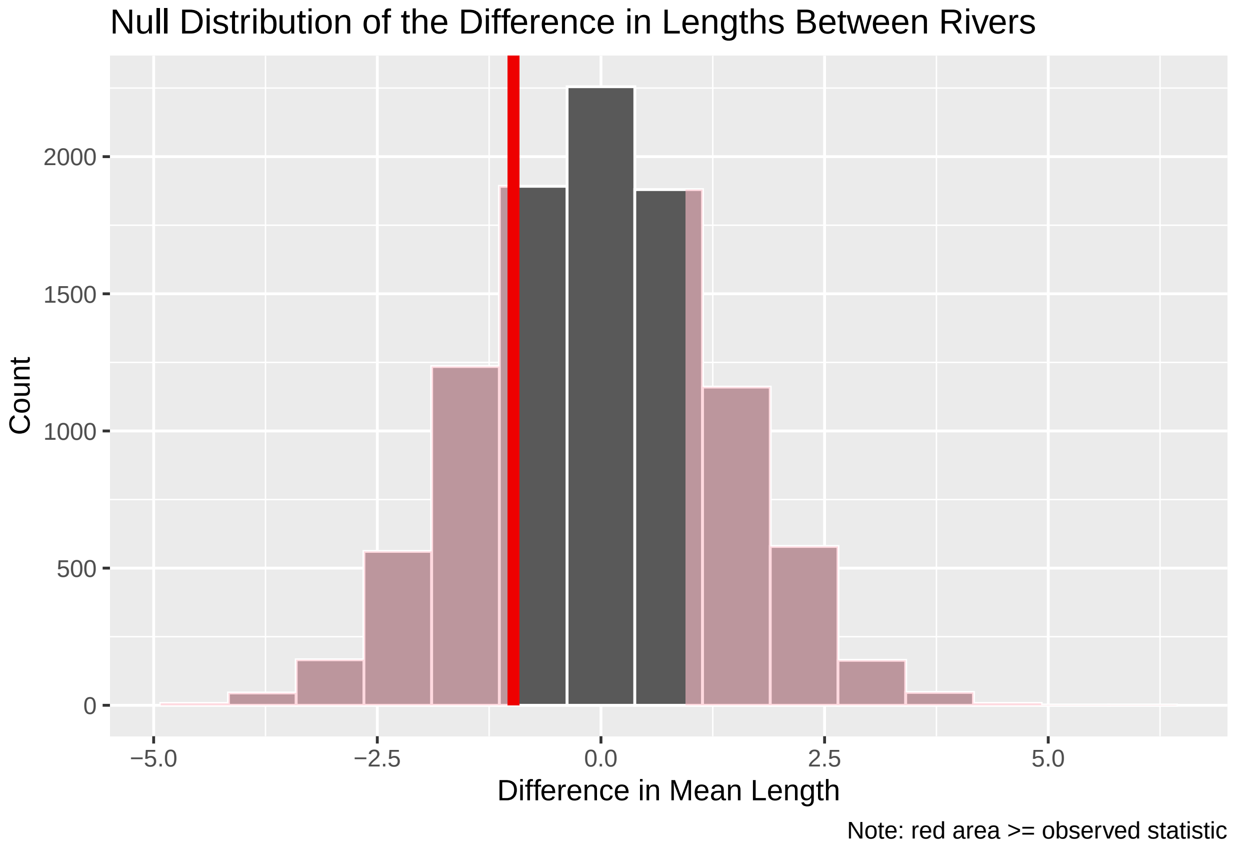
Does the average length of bass differ between the two rivers?

```
set.seed(1089336)
diff_length <- mercury_bass %>%
  group_by(riverName) %>%
  summarize(mean_length = mean(length)) %>%
  summarize(diff_length = mean_length[riverName == "Lumber"] -
              mean_length[riverName == "Waccamaw"]) %>%
  pull()

null_dist_3 <- mercury_bass %>%
  specify(response = length, explanatory = riverName) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means")
```

```
## Warning: The statistic is based on a difference or ratio; by default, for
## difference-based statistics, the explanatory variable is subtracted in the order
## "Lumber" - "Waccamaw", or divided in the order "Lumber" / "Waccamaw" for ratio-
## based statistics. To specify this order yourself, supply `order = c("Lumber",
## "Waccamaw")` to the calculate() function.
```

```
visualize(null_dist_3) +
  shade_p_value(obs_stat = diff_length, direction = "both") +
  labs(title = "Null Distribution of the Difference in Lengths Between Rivers",
       x = "Difference in Mean Length", y = "Count",
       caption = "Note: red area >= observed statistic")
```


Null Distribution of the Difference in Lengths Between Rivers

Note: red area >= observed statistic

```
#Note: while I am aware the question didn't ask us to calculate the p-value, I
#would hate to not do so after doing all this work
null_dist_3 %>%
  get_p_value(obs_stat = diff_length, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.473
```

Based on this visualization (ignoring the p-value, since I wasn't actually asked to calculate that), it does not seem like there is much difference between the length of bass found in the two rivers beyond random chance.