CSC100/CSC200 Lab #9: Hypothesis Testing

Dominick Bello

Fall 2022 | Data Science for the World

Please complete this notebook by filling in the cells provided. When you're done:

1. Remember to put your name in the header at the top of this notebook where it says `author`.
2. Select `Knit (Knit to Word)` from the toolbar menu.
3. Read that file! If any of your lines are too long and get cut off, we won't be able to see them, so break them up into multiple lines and knit again.
4. Save that Word document as a PDF file.
5. Submit BOTH this `.Rmd` file and the **PDF** file you generated to Gradescope. Some questions are autograded and you may improve your score on the tests given by resubmitting your work as many times as you like up to the deadline.
6. **Passing the automatic tests given does not guarantee full credit on any question.** The tests are provided to help catch some common mistakes, but it is *your* responsibility to answer the questions correctly.

If you are having trouble submitting, ask your lab instructor for help.

This lab assignment is due **October 31 at 9:00AM**.

Reading:

- Chapter 8 textbook

Run the cell below to prepare the notebook.

U.S. Supreme Court, 1965: Swain vs. Alabama

Swain v. Alabama was a US Supreme Court case decided in 1965. A black man, Swain, was accused of raping a white woman, and had been convicted by an all-white jury. The jury was selected from a panel of 100 people that contained only 8 black people and 92 people of other ethnicities. That panel was supposed to be a random sample from the eligible population of Talladega County, which was 26% black and 74% other ethnicities (and, by law, all male). We will assume there were 16,000 people in all of Talladega County eligible to serve on the jury.

Swain's lawyers argued that this reflected a bias in the jury panel selection process. A five-justice majority of the Supreme Court disagreed, asserting that,

> "The overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negros."

Is 8 enough smaller than 26 to indicate that the jury panel wasn't just a random sample from the eligible population?

It *could* have happened by chance, so this may seem like a judgement call. Indeed, five Supreme Court justices apparently thought it was a matter of judgement.

In fact, with a computer simulation, we can answer that is is very unlikely to see a panel of 8 randomly selected from this population.

The tibble `jury` gives the data:

```
jury <- tribble(~ethnicity, ~eligible, ~panel,
                "Black", 4160, 8,
                "Other", 11840, 92)
jury

## # A tibble: 2 × 3
##   ethnicity eligible panel
##   <chr>        <dbl> <dbl>
## 1 Black         4160     8
## 2 Other        11840    92
```

Part I: Visualization. As always, a good place to begin is visualization. Because `ethnicity` is a categorical variable, a bar chart would be a helpful visualization to better understand the problem.

Question 1. Something isn't right about the tibble `jury`. Why is a bar chart visualization not a good idea with the data we have right now? To help you figure it out, try drawing the bar chart by hand and see what happens.

'Other' is a non-descriptive categorical variable that represents a large majority of the data.

Question 2. Fix the problem by creating a new tibble called `jury_prop` that is a copy of `jury` but contains the adjusted values. Your tibble should have the same three variables: `ethnicity`, `eligible`, and `panel`. The provided test will check if you got it right.

```
jury_prop <- tribble(~ethnicity, ~eligible, ~panel,
                "Black", .26, .08,
                "Other", .74, .92)

jury_prop

## # A tibble: 2 × 3
##   ethnicity eligible panel
##   <chr>        <dbl> <dbl>
## 1 Black         0.26  0.08
## 2 Other         0.74  0.92

. = ottr::check("tests/visualization_q2.R")

## All tests passed!
```

Question 3. We are ready to visualize – but the form of the tibble `jury_prop` is not suitable for direct application of `ggplot`! (why?) Apply a pivot longer transformation to `jury_prop` so that there are three variables: `ethnicity`, `group`, and `count`. Assign the resulting tibble to the name `jury_prop_long`.

```
jury_prop_long <- jury_prop %>%
  pivot_longer(!ethnicity, names_to = "group", values_to = "count")



jury_prop_long
```

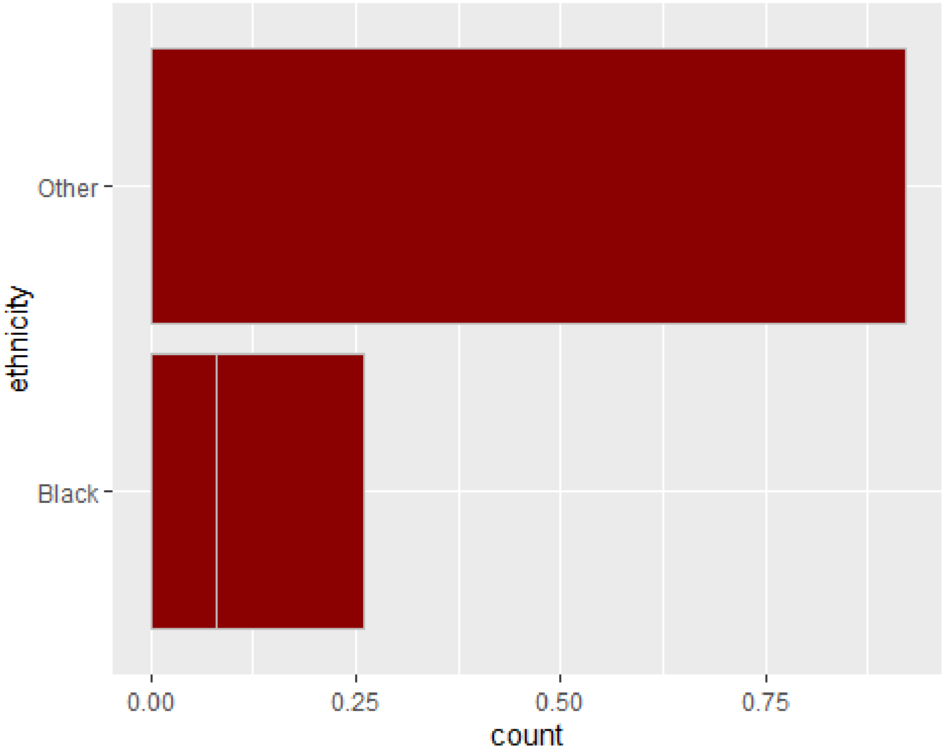
```
## # A tibble: 4 × 3
##   ethnicity group    count
##   <chr>     <chr>    <dbl>
## 1 Black     eligible  0.26
## 2 Black     panel     0.08
## 3 Other     eligible  0.74
## 4 Other     panel     0.92
```

```
. = ottr::check("tests/visualization_q3.R")
```

```
## All tests passed!
```

Question 4. Create a horizontal bar chart showing both the proportion of ethnicities in the panel and in the eligible population. You should use a bar geom and a coordinate system. You should also consider using a positional adjustment like "dodge." Check with a neighbor or lab instructor to make sure you got the plot right.

```
 ggplot(jury_prop_long, aes(x = count, y = ethnicity)) +
  geom_col(position="dodge", color = "gray", fill = "darkred")
```



Part II: Hypothesis Testing. To test the hypothesis that the actual panel was a random sample from the eligible population, we first write down a *null hypothesis*. We will imagine (via computer simulation) what the data would typically look like if this hypothesis were true. If the actual data don't look like that, we'll reject the null hypothesis.

Our null hypothesis is as follows:

> **Null hypothesis:** "The actual panel in Swain's trial was a random sample from the eligible population in Talladega County."

Now, imagine drawing a random sample of 100 people from among the eligible jurors. This is one panel we could see *if the null hypothesis were true*. We can simulate drawing one sample "panel" from the population of eligible jurors in Talladega by calling the function `rmultinom()` we saw in lecture:

```
eligible_distribution <- jury_prop %>% pull(eligible)
rmultinom(n = 1, size = 100, prob = eligible_distribution)

##      [,1]
## [1,]   32
## [2,]   68
```

The first element in this vector contains the number of black people in this sample panel; likewise, the second element for the number of other ethnicities.

Question 1. For a distribution of ethnicities in our sample, we are more interested in the *proportion* of resulting ethnicities that appear in the panel. Write a function sample_proportions() that takes any distribution (e.g., eligible_distribution) as an argument and returns a vector containing the proportions of ethnicities that appear in the sample panel of 100 people.

```
sample_proportions <- function(distribution) {
distribution <- jury_prop %>% pull(eligible)
  return(c(rmultinom(n = 1, size = 100, prob = distribution)/100) )
}
sample_proportions(eligible_distribution)

## [1] 0.27 0.73

. = ottr::check("tests/hypothesis_q1.R")

## Test hypothesis_q1 - 1 failed:
## 
## between(res[1], 0.01, 0.3) is not TRUE
## 
##   `actual`:   FALSE
##   `expected`: TRUE
```

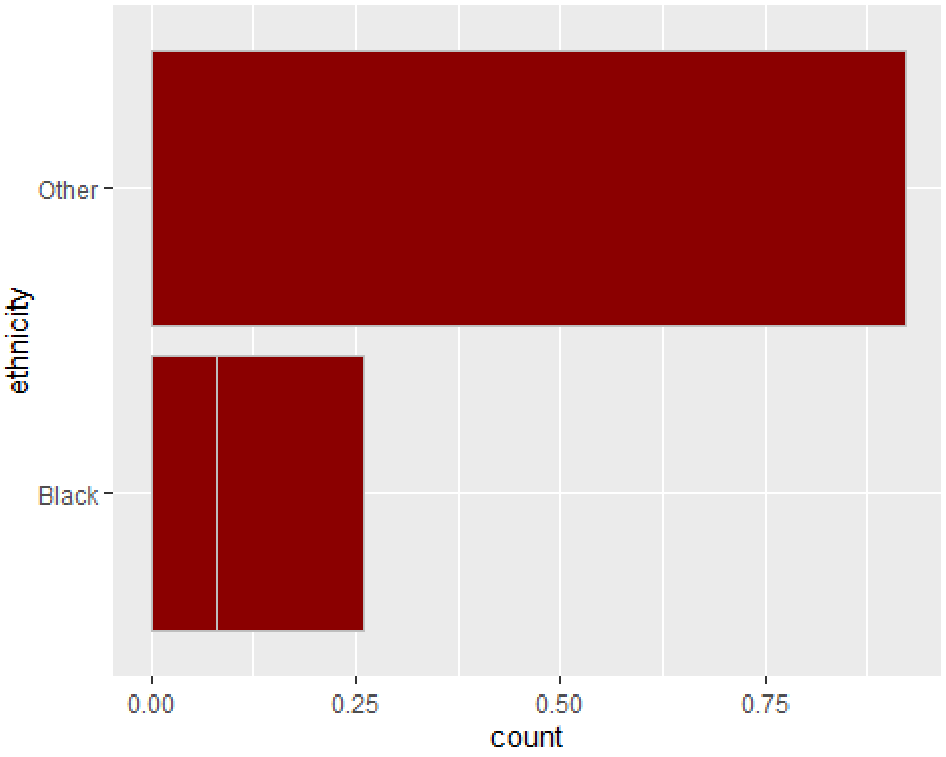
Question 2. Call sample_proportions() to create one vector called one_sample_of_100 that represents one sample of 100 people from among the eligible jurors. This is one panel we could see *if the null hypothesis were true*.

```
one_sample_of_100 <- sample_proportions(eligible_distribution)
one_sample_of_100

## [1] 0.26 0.74

. = ottr::check("tests/hypothesis_q2.R")

## All tests passed!
```

Question 3. Recreate the horizontal bar chart, this time also displaying the proportions of ethnicities in this sample. That means *three* proportions should be presented in a single bar chart: (1) the proportions of ethnicities in the eligible population; (2) in the actual panel in Swain's case; (3) and in the sample one_sample_of_100.

HINT: You will need to repeat some steps from the above questions; most of the code you need you have already written.

```
 ggplot(jury_prop_long, aes(x = count, y = ethnicity, fill = group)) +
  geom_col(position="dodge", color = "gray", fill = "darkred")
```



Does the panel look like it could have come from this process?

To answer that question, we'll need to sample many times, not just once. And we'll need to summarize each sample with a number (a "test statistic"). We want the number to generally look one way if the null hypothesis is true, and some other way if it's not.

A useful test statistic in cases like this is the *total variation distance* (TVD) between the distribution of ethnicities in the sample and the distribution of ethnicities in the eligible population. Intuitively, this distance should be typically small if the null hypothesis is true, because many samples will have similar proportions of ethnicities as the population from which they're taken.

We can actually compute the TVD visually from the above bar chart you made. For each category ("Black" and "Other"), find the absolute difference between the lengths of the bars. Add up those absolute differences, and divide by 2.

Question 4. Without using any code, estimate the TVD between the distribution of ethnicities in the sample and the distribution of ethnicities in the eligible population using the above bar chart. Then estimate the TVD between the distribution of ethnicities in the *actual panel* and the distribution of ethnicities in the eligible population.

Note which one is bigger. Check with a neighbor or a TA to verify your answer.

```r
rough_tvd_between_sample_and_population <- .05
rough_tvd_between_panel_and_population <- 0.2

. = ottr::check("tests/hypothesis_q4.R")
```

```
## All tests passed!
```

Question 5. Write a function called `tvd_from_eligible_population()`. It should take one argument: a vector of proportions of ethnicities. The first element in the vector is the proportion of black people, and the second element is the proportion of others. The function should return the TVD between that distribution of ethnicities and the distribution in the eligible population.

```
tvd_from_eligible_population <- function(proportions) {
  return(sum(abs(proportions - eligible_distribution)) / 2)
}

# An example call to your function. This computes the
# TVD you estimated in question 4.
print(paste("TVD between eligible population and a random sample:",
tvd_from_eligible_population(one_sample_of_100)))

## [1] "TVD between eligible population and a random sample: 0"

# ...and this computes the other TVD you estimated.
panel_distribution <- jury_prop %>% pull(panel)
print(paste("TVD between eligible population and the actual panel:",
tvd_from_eligible_population(panel_distribution)))

## [1] "TVD between eligible population and the actual panel: 0.18"

. = ottr::check("tests/hypothesis_q5.R")

## All tests passed!
```

Now you have all the ingredients for running a hypothesis test:

1. You can generate data you would see if the null hypothesis were true. You did that once in question 2 with `one_sample_of_100`.
2. You can compute a test statistic on each simulated "panel" by calling `tvd_from_eligible_population`.

Question 6. Write a function called `one_simulated_tvd()`. The function takes no arguments. It should generate a "sample panel" under the null hypothesis, compute the test statistic, and then return it.

```
one_simulated_tvd <- function() {
  sample_panel <- sample_proportions(eligible_distribution)
  return(tvd_from_eligible_population(sample_panel))
}

one_simulated_tvd()  # an example call

## [1] 0.04

. = ottr::check("tests/hypothesis_q6.R")
```

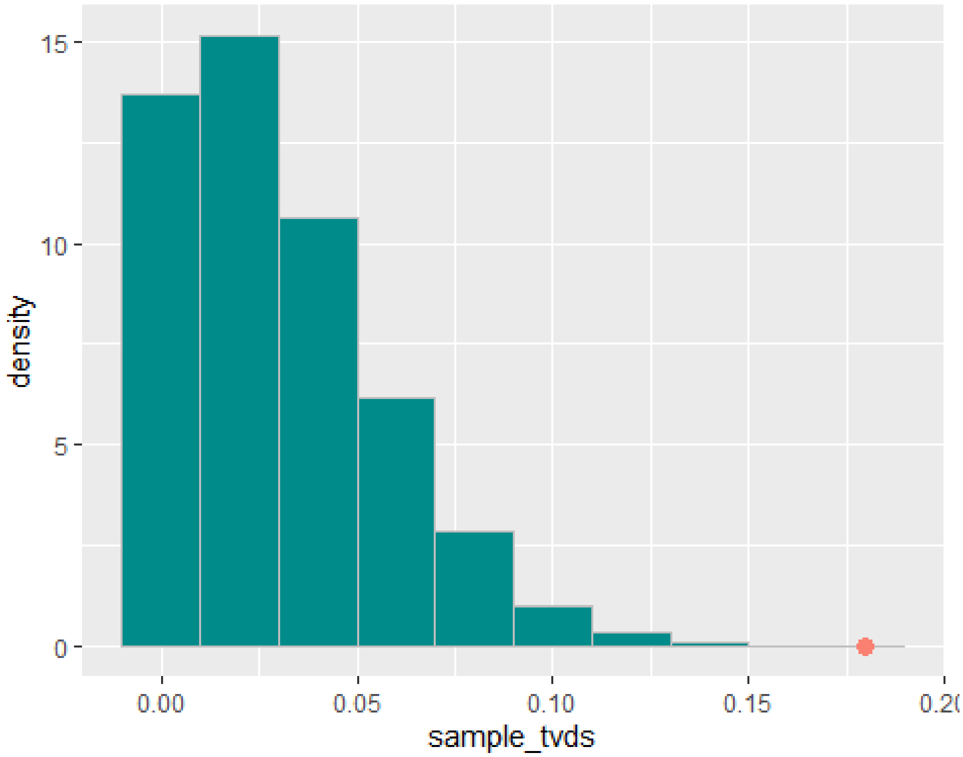
```
## All tests passed!
```

Question 7. Using `replicate()`, run the simulation **10,000** times to produce **10,000** test statistics. Assign the results to a vector called `sample_tvds`.

```
sample_tvds <- replicate(n = 10000, one_simulated_tvd())

. = ottr::check("tests/hypothesis_q7.R")

## All tests passed!
```

Run the cell below to see the results of your simulation augmented with an orange dot that shows the TVD between the actual panel and the eligible population (which you estimated in question 4).

```
ggplot(tibble(sample_tvds)) +
  geom_histogram(aes(x = sample_tvds, y = ..density..), bins = 10,
                 fill = "darkcyan", color = 'gray') +
  geom_point(aes(x = tvd_from_eligible_population(panel_distribution), y =
0),
             size = 3, color = "salmon")
```



Question 8. Do we have evidence for or against the null hypothesis, or not much evidence either way? Discuss with a neighbor and/or a lab instructor.

We have evidence that this goes against the null hypotheseis.

Yahoo – you are done with Lab #9! Time to submit.