CSC200 Project: Inference and Capital Punishment

Caroline Hall

Fall 2021 | Data Science for the World | October 18, 2021

This project investigates the relationship between murder and capital punishment (the death penalty) in the United States. By the end of the project, you should know how to:

1. Test whether observed data appears to be a random sample from a distribution
2. Analyze a natural experiment
3. Implement and interpret a sign test
4. Create a function to run a general hypothesis test
5. Generate and analyze visualizations, and then draw conclusions from them

Housekeeping

Discussion Board

While collaboration is encouraged, sharing answers is never okay. In particular, posting code or other assignment answers publicly on the BB Discussion Board (or elsewhere) is academic dishonesty. It will result in a reduced project grade at a minimum. If you wish to ask a question that involves code, you *must* reach out to a TA or an instructor for help.

Checkpoint and Due Date

Part of the project will be due early. Parts 1 and 2 of the project (out of 5) are due **Monday, November 8 at 11:59PM**. Unlike the final submission, this early checkpoint will be graded for completion. That is, only the public tests for these parts must be passing to clear the checkpoint. It will be worth approximately 10% of the total project grade. Simply submit your partially-completed notebook as you would submit any other notebook.

The entire project (parts 1, 2, 3, 4, and 5) will be due **Monday, December 6 at 11:59PM**.

On to the project!

Run the cell below to prepare the automatic tests. The automated tests for this project **definitely do not** catch all possible errors; they're designed to help you avoid some common mistakes. Merely passing the tests does not guarantee full credit on any question.

Part I: Murder Rates

Punishment for crime has many philosophical justifications. An important one is that fear of punishment may *deter* people from committing crimes.

In the United States, some jurisdictions execute some people who are convicted of particularly serious crimes, like murder. This punishment is called the *death penalty* or *capital punishment*. The death penalty is controversial, and deterrence has been one focal point of the debate. There are other reasons to support or oppose the death penalty, but in this project we'll focus on deterrence.

The key question about deterrence is:

> Does instituting a death penalty for murder actually reduce the number of murders?

You might have a strong intuition in one direction, but the evidence turns out to be surprisingly complex. Different sides have argued that the death penalty has no deterrent effect and that each execution prevents 8 murders, all using statistical arguments! We will try to form our own conclusion.

Here is a road map for this project:

1. In the rest of this section, we'll investigate the main dataset we'll be using.
2. In section 2, we'll see how to test null hypotheses like this: "For this set of U.S. states, the murder rate was equally likely to go up or down each year."
3. In section 3, we'll apply a similar test to see whether U.S. states that suddenly ended or reinstated the death penalty were more likely to see murder rates increase than decrease.
4. In section 4, we will run some more tests to further claims developed in previous sections.
5. In section 5, we'll try to answer our question about deterrence using a visualization rather than a formal hypothesis test.

Dataset

The main data source for this project comes from a paper by three researchers, Dezhbakhsh, Rubin, and Shepherd. The dataset contains rates of various violent crimes for every year 1960-2003 (44 years) in every US state. The researchers compiled their data from the FBI's Uniform Crime Reports.

Since crimes are committed by people, not states, we need to account for the number of people in each state when we're looking at state-level data. Murder rates are calculated as follows:

$$\text{murder rate for state X in year Y} = \frac{\text{number of murders in state X in year Y}}{\text{population in state X in year Y}} * 100000$$

Murder is rare, so we multiply by 100,000 just to avoid dealing with tiny numbers.

Question 1.1. Load in this dataset from `data/crime_rates.csv` and select only the columns `State`, `Year`, `Population`, and `Murder Rate`. Assign the loaded tibble to the name `murder_rates`.

```r
murder_rates <- read_csv("data/crime_rates.csv") %>%
  select(State, Year, Population, `Murder Rate`)
```

```
## Rows: 2200 Columns: 12
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (1): State
## dbl (11): Year, Population, Violent Crime Rate, Murder Rate, Forcible Rape
R...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
murder_rates
```

```
## # A tibble: 2,200 × 4
##    State   Year Population `Murder Rate`
##    <chr>  <dbl>      <dbl>         <dbl>
##  1 Alaska  1960     226167         10.2
##  2 Alaska  1961     234000         11.5
##  3 Alaska  1962     246000          4.5
##  4 Alaska  1963     248000          6.5
##  5 Alaska  1964     250000         10.4
##  6 Alaska  1965     253000          6.30
##  7 Alaska  1966     272000         12.9
##  8 Alaska  1967     272000          9.60
##  9 Alaska  1968     277000         10.5
## 10 Alaska  1969     282000         10.6
## # … with 2,190 more rows
```

```
. = ottr::check("tests/part1_q1.R")
```

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

So far, these data aren't enough to demonstrate an *association* between the existence of the death penalty in a state in a year and the murder rate in that state and year.

Question 1.2. What additional information will we need before we can check for that association?

The additional information we need is whether or not the death penalty exists for each state and year.

Murder rates vary over time, and different states exhibit different trends. The rates in some states change dramatically from year to year, while others are quite stable. Let's plot a couple, just to see the variety.

Question 1.3. Filter `murder_rates` to include murder rate data *only* for New Mexico and Minnesota. Call the resulting tibble `nm_mn`.

```
nm_mn <- filter(murder_rates, State == "New Mexico" | State == "Minnesota")
nm_mn
```

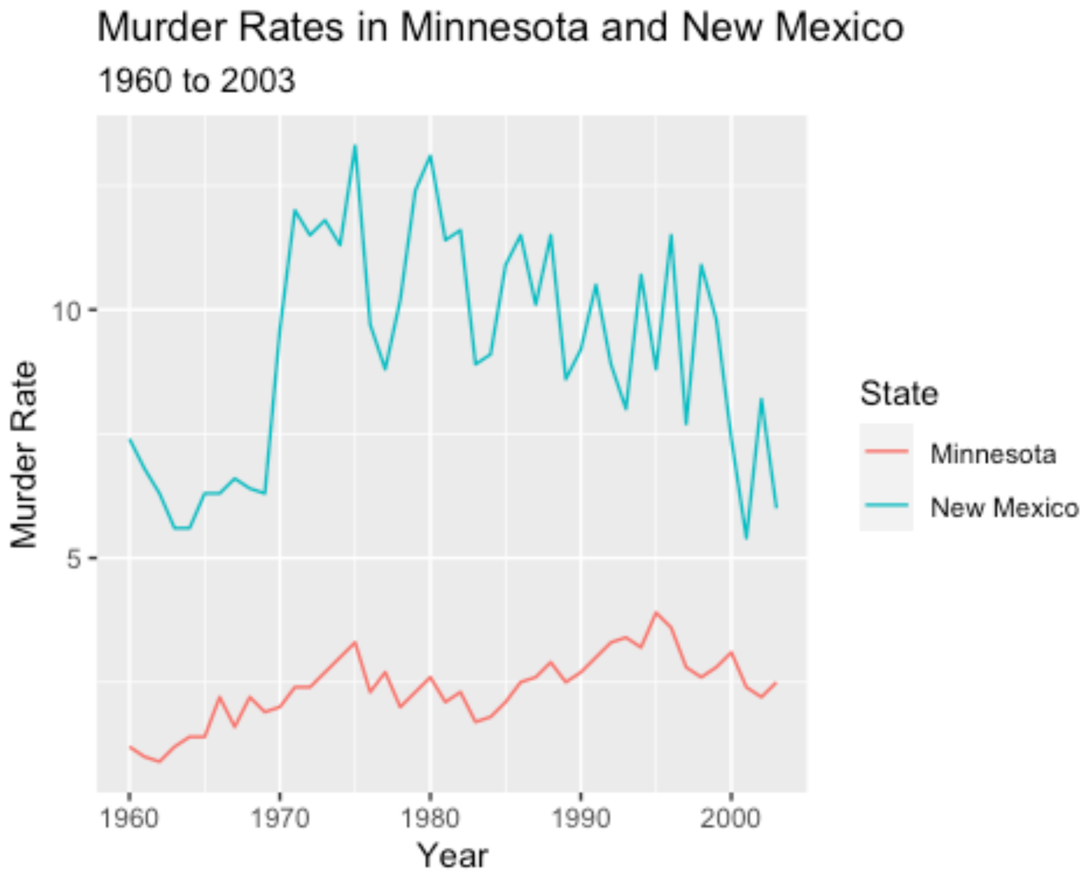
```
## # A tibble: 88 × 4
##    State      Year Population `Murder Rate`
##    <chr>     <dbl>      <dbl>         <dbl>
##  1 Minnesota  1960    3413864         1.20
##  2 Minnesota  1961    3470000         1
##  3 Minnesota  1962    3475000         0.900
##  4 Minnesota  1963    3500000         1.20
##  5 Minnesota  1964    3521000         1.40
##  6 Minnesota  1965    3554000         1.40
##  7 Minnesota  1966    3576000         2.20
##  8 Minnesota  1967    3582000         1.60
##  9 Minnesota  1968    3646000         2.20
## 10 Minnesota  1969    3700000         1.90
## # … with 78 more rows
```

```
. = ottr::check("tests/part1_q3.R")
```

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

Question 1.4. Draw a line plot using `ggplot2` with years on the horizontal axis and murder rates on the vertical axis. Two lines should be shown in a *single* plot: one for New Mexico murder rates and one for Minnesota murder rates.

```
ggplot(nm_mn, aes(x=Year, y=`Murder Rate`, color=State)) +
  geom_line() +
  labs(title = "Murder Rates in Minnesota and New Mexico",
       subtitle = "1960 to 2003")
```



Question 1.5. How many more *people* were murdered in Florida in 1988 than in 1987? Assign `fl_change` to the answer.

Hint: Consider using the formula in the beginning of the section to answer this question. You will also need to use `dplyr` code.

```
fl_change <- murder_rates %>%
  filter(State=="Florida", Year=="1987" | Year=="1988") %>%
  mutate(number_of_murders = (`Murder Rate`*Population)/100000) %>%
  pull(number_of_murders) %>%
  diff() %>%
  round()
fl_change

## [1] 40

. = ottr::check("tests/part1_q5.R")

## All tests passed!
```

Make sure your answer looks reasonable after carefully reading the question.

Part II: Changes in Murder Rates

Murder rates vary considerably across states and years, presumably due to the vast array of differences among states and across U.S. history. Rather than attempting to analyze rates themselves, here we will restrict our analysis to whether or not murder rates increased or decreased over certain time spans. We will not concern ourselves with how much rates increased or decreased; only the direction of the change - *whether* they increased or decreased.

The `diff` function takes a vector of values and computes the differences between adjacent items of a vector. However, we may wish to compute the difference between items that are two positions apart. For example, given a 5-element vector, we may want:

```
[item 2 - item 0 , item 3 - item 1 , item 4 - item 2]
```

Here is an example:

```
diff(c(1, 10, 100, 1000, 10000), 2)

## [1]   99  990 9900
```

Question 2.1. Implement the function `two_year_changes()` that takes a *vector* of murder rates for a state, ordered by increasing year. For all two-year periods (e.g., from 1960 to 1962), it computes the number of increases minus the number of decreases, and returns the result.

For example:

```
r <- c(10, 7, 12, 9, 13, 9, 11)
r

## [1] 10  7 12  9 13  9 11
```

`r` contains three increases (10 to 12, 7 to 9, and 12 to 13), one decrease (13 to 11), and one change that is neither an increase or decrease (9 to 9). Therefore, `two_year_changes(r)` would return 2, the difference between 3 increases and 1 decrease.

```
two_year_changes<- function(x){
  diff(x, 2) %>%
    sign() %>%
    sum()
}

two_year_changes(r) # an example call

## [1] 2

. = ottr::check("tests/part2_q1.R")

## All tests passed!
```

We can use two_year_changes() to summarize whether rates are mostly increasing or decreasing over time for some state or group of states. Let's see how it varies across all 50 US states.

Question 2.2. Assign changes_by_state to a tibble with one row per state that has two columns: the State name and Total differences which gives the two_year_changes statistic computed across all years in the dataset for that state. The first 2 rows of the tibble should look like this:

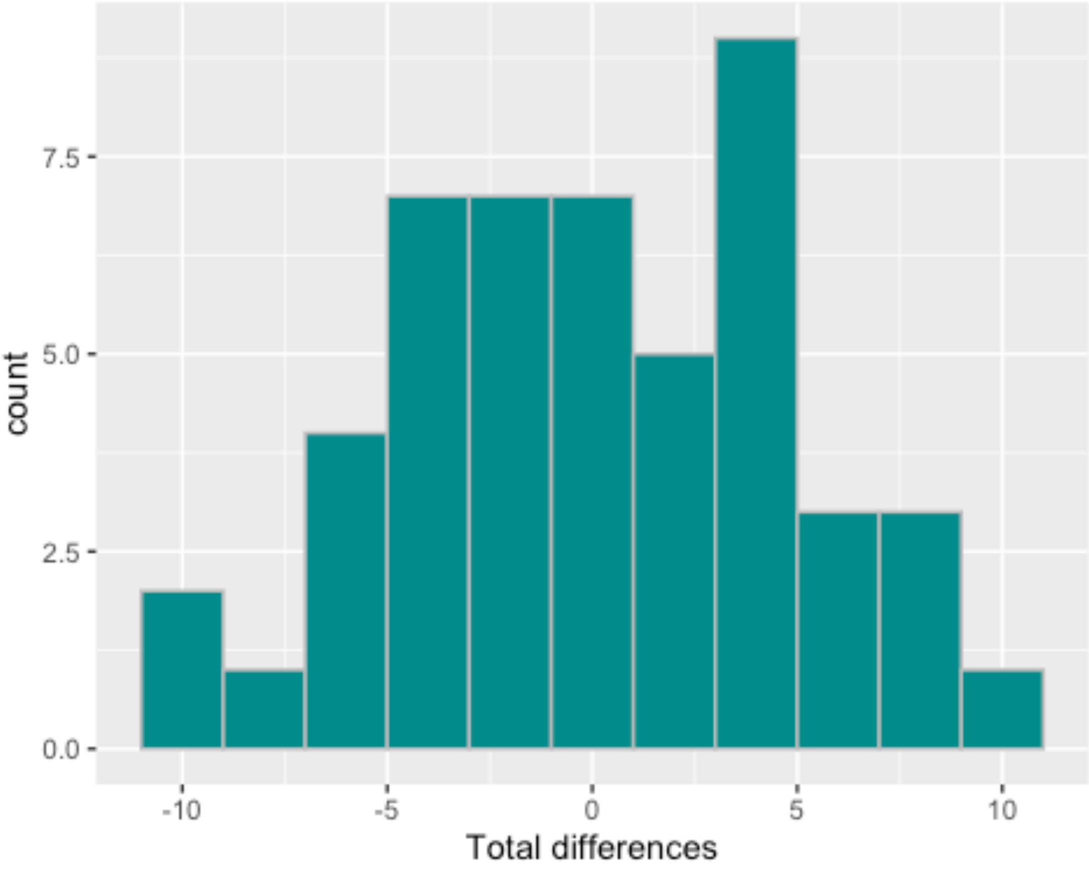
State	Total differences
Alabama	-6
Alaska	-5
…	…

To accomplish this, you will need to group by and summarize *within* each group. We saw how to compute within groups in Homework #4. The automatic test will check if your resulting tibble looks right.

```
changes_by_state<- murder_rates %>%
  group_by(State) %>%
  summarize(`Total differences` = two_year_changes(`Murder Rate`)) %>%
  ungroup()
changes_by_state

## # A tibble: 50 × 2
##    State       `Total differences`
##    <chr>                     <dbl>
##  1 Alabama                      -6
##  2 Alaska                       -5
##  3 Arizona                       1
##  4 Arkansas                     -1
##  5 California                   17
##  6 Colorado                     -4
##  7 Connecticut                   4
##  8 Delaware                     -3
##  9 Florida                      -6
## 10 Georgia                      -3
## # … with 40 more rows

. = ottr::check("tests/part2_q2.R")

## All tests passed!
```

Question 2.3. Generate a histogram of the two-year changes for the states using your changes_by_state tibble. We suggest using the following bins:

```
changes_by_state_bins <- seq(-11, 12, 2)

ggplot(changes_by_state, aes(x=`Total differences`)) +
  geom_histogram(fill = "darkcyan",
                 color = "gray",
                 breaks = changes_by_state_bins)
```



Some states have more increases than decreases (a positive number), while some have more decreases than increases (a negative number).

Question 2.4. Assign the name `total_changes` to the total increases minus the total decreases for all two-year periods and all states in the dataset `changes_by_state`.

```
total_changes<- changes_by_state %>%
  pull(`Total differences`) %>%
  sum()
total_changes

## [1] 45

. = ottr::check("tests/part2_q4.R")

## All tests passed!
```

A student remarks about your findings: “Murder rates tend to go up across two-year periods. What a time to be alive…”

Haven taken CSC100/CSC200, you reply, “Not so fast! Even if murder rates moved up and down uniformly at random, there would be some difference between the increases and decreases. There were a lot of states and a lot of years, so there were many chances for changes to happen. Perhaps this difference we observed is a typical value when so many changes are observed if the state murder rates increase and decrease at random!”

Question 2.5. Set `num_changes` to the number of different two-year periods in the entire data set that *could* result in a change of a state’s murder rate. Include both those periods where a change occurred and the periods where a state’s rate happened to stay the same.

For example, 1968 to 1970 of New Mexico would count as one distinct two-year period.

```r
num_changes <- (nrow(murder_rates)/2) - 2
num_changes

## [1] 1098

. = ottr::check("tests/part2_q5.R")

## All tests passed!
```

We now have enough information to perform a hypothesis test.

> **Null Hypothesis**: State murder rates increase and decrease over two-year periods as if “increase” or “decrease” were sampled at random from a uniform distribution, like a fair coin flip.

Since it’s possible that murder rates are more likely to go up or more likely to go down, our alternative hypothesis should account for both cases:

> **Alternative Hypothesis**: State murder rates are *either* more likely or less likely to increase than decrease over two-year periods.

Note: These changes in murder rates are not random samples from any population. They describe all murders in all states over all recent years. However, we can imagine that history could have been different, and that the observed changes are the values observed in only one possible world: the one that happened to occur. In this sense, we can evaluate whether the observed “total increases minus total decreases” is consistent with a hypothesis that increases and decreases are drawn at random from a uniform distribution.

Question 2.6 Given these null and alternative hypotheses, define a good test statistic by writing a function `test_stat()` that implements the test statistic you have chosen.

Important requirements for your test statistic: Choose a test statistic for which *only* large positive values are evidence in favor of the alternative hypothesis, and any other values are insufficient evidence. Your test statistic should depend only on *whether* murder rates increased or decreased, not on the *size* of any change.

```r
test_stat<- function(x){
  y<- abs(sum(x))
  return(y)
}
```

The cell below returns a vector after sampling “increases” and “decreases” at random from a uniform distribution 100 times.

```
my_sample <- sample(c(1, -1), size=100, replace = TRUE)
```

And here is what the test statistic for the sample looks like after running it through your `test_stat()` function:

```
test_stat(my_sample)

## [1] 0
```

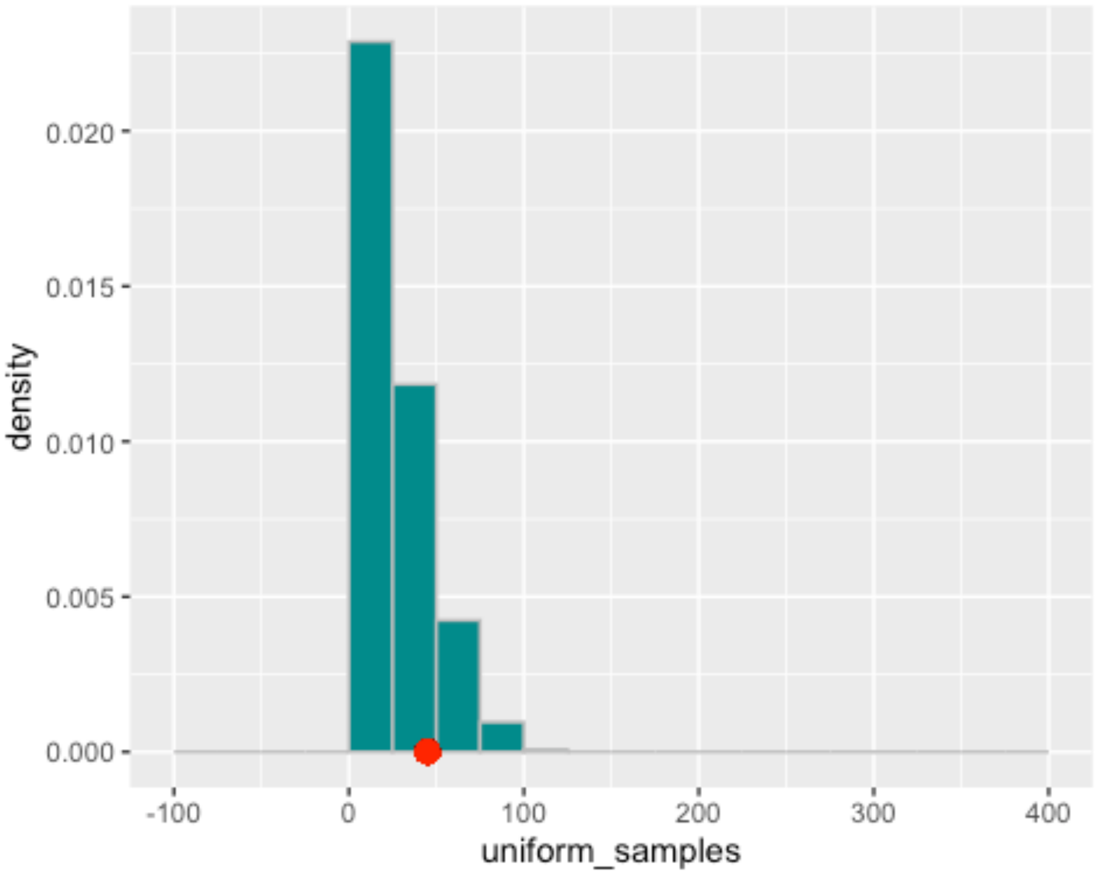
Note: You may be wondering how we can say whether the murder rate was more likely to increase *or* more likely to decrease when the test statistic does not care about the size of the change. We can tell by looking at the size of the “total increases minus total decreases” (as we did in Question 2.4) before computing the test statistic. For instance, if we find that value to be *negative* and there is significant evidence in favor of the *alternative hypothesis*, then we can say that the murder rate is indeed more likely to *decrease*.

Question 2.7. Complete the simulation below, which samples `num_changes` increases/decreases at random many times. Your job is to fill in the function `simulate_under_null()`, which simulates a single sample under the null hypothesis.

```
simulate_under_null <- function(x) {
  # Simulates some number changing several times, with an equal
  # chance to increase or decrease.
  # Returns the value of your test statistic for these simulated
  # changes.

  num_chances_to_change<- test_stat(sample(c(1, -1), size=x, replace = TRUE))
  return(num_chances_to_change)
}



uniform_samples <- replicate(n = 500, simulate_under_null(num_changes))
```

The following cell forms a sampling histogram from the samples you generated in `uniform_samples`. It also annotates the histogram with where `total_changes` is located.

```
ggplot(tibble(uniform_samples)) +
  geom_histogram(aes(x = uniform_samples, y = ..density..), fill =
"darkcyan",
                 color="gray", breaks=seq(-100, 400, 25), closed="left") +
  geom_point(aes(x = total_changes, y = 0), size = 3, color = "red")
```



Question 2.8. Looking at this histogram, draw a conclusion about whether murder rates basically increase as often as they decrease. Computing a formal p-value is not necessary for answering this question.

We can see from this histogram that murder rates tend to increase about as often as they decrease. Based on the location of our test statistic from the murder rates data on the histogram, we conclude that we do not have enough evidence to reject the null hypothesis that state murder rates increase and decrease over two-year periods as if “increase” or “decrease” were sampled at random from a uniform distribution.

You reached the project checkpoint! Congrats!

Remember to submit your **partially completed .Rmd notebook file** to Gradescope by the Stage I deadline if you have not done so already. The checkpoint is graded only on successful completion of the public tests given; accuracy of your Part I and Part II work will be assessed after the final deadline.

HOWEVER, do **NOT** continue on to Part III until you are confident in your answers for Parts I and II and have confirmed that all the automatic tests up to this point are passing!

Part III: The death penalty.

Some US states have the death penalty, others don't, and laws have changed over time. In addition to changes in murder rates, we will also consider whether the death penalty was in force in each state and each year.

Using this information, we would like to investigate how the death penalty affects the murder rate of a state.

A natural experiment

In order to attempt to investigate the causal relationship between the death penalty and murder rates, we're going to take advantage of a *natural experiment*. A natural experiment happens when something other than experimental design applies a treatment to one group and not to another (control) group, and we can expect that the treatment and control groups don't have any other systematic differences.

Our natural experiment is the Supreme Court decision in 1972 called *Furman v. Georgia* that banned the death penalty throughout the US. Suddenly, many states went from having the death penalty to not having the death penalty.

As a first step, let's see how murder rates changed before and after the court decision. We'll define the test as follows:

> **Population:** All the states that had the death penalty before the 1972 abolition. This includes every US state **other than** Alaska, Hawaii, Maine, Michigan, Wisconsin, and Minnesota.
>
> **Treatment group:** The states in the above population, in the year after 1972.
>
> **Control group:** The states in the above population, in the year before 1972.
>
> **Null hypothesis:** Each state's murder rate was equally likely to be higher or lower in the treatment period than in the control period. In other words, whether the murder rate increased or decreased in each state was like the flip of a fair coin.
>
> **Alternative hypothesis:** The murder rate was more likely to increase *or* more likely to decrease.

Here we have defined a vector `non_death_penalty_states` containing those US states that did not have the death penalty in 1971. We also define a convenience function `had_death_penalty_in_1971()` that returns `TRUE` if the argument is the name of a state that had the death penalty in 1971 and `FALSE` otherwise.

```r
non_death_penalty_states <- c('Alaska', 'Hawaii', 'Maine', 'Michigan',
                              'Wisconsin', 'Minnesota')

had_death_penalty_in_1971 <- function(state) {
  return(!(state %in% non_death_penalty_states))
}
```

Question 3.1. Add a new logical variable to murder_rates called death_penalty_in_1971 that is TRUE if the corresponding state for that row had the death penalty in 1971 and FALSE otherwise. Call the resulting tibble murder_rates_death_penalty.

```
murder_rates_death_penalty<- murder_rates %>%
  mutate(death_penalty_in_1971 = had_death_penalty_in_1971(State))
murder_rates_death_penalty

## # A tibble: 2,200 × 5
##    State   Year Population `Murder Rate` death_penalty_in_1971
##    <chr>  <dbl>      <dbl>         <dbl> <lgl>
##  1 Alaska  1960     226167         10.2  FALSE
##  2 Alaska  1961     234000         11.5  FALSE
##  3 Alaska  1962     246000          4.5  FALSE
##  4 Alaska  1963     248000          6.5  FALSE
##  5 Alaska  1964     250000         10.4  FALSE
##  6 Alaska  1965     253000          6.30 FALSE
##  7 Alaska  1966     272000         12.9  FALSE
##  8 Alaska  1967     272000          9.60 FALSE
##  9 Alaska  1968     277000         10.5  FALSE
## 10 Alaska  1969     282000         10.6  FALSE
## # … with 2,190 more rows

. = ottr::check("tests/part3_q1.R")

## All tests passed!
```

Question 3.2. How many states had the death penalty implemented in 1971? Use dplyr code to answer this. Assign your answer to the name num_death_penalty_states.

```
num_death_penalty_states<- murder_rates_death_penalty %>%
  group_by(State) %>%
  summarize(yes_or_no_death_penalty = sign(sum(death_penalty_in_1971))) %>%
  pull(yes_or_no_death_penalty) %>%
  sum()
num_death_penalty_states

## [1] 44

. = ottr::check("tests/part3_q2.R")

## All tests passed!
```

Question 3.3. Assign murder_rates_death_penalty_active to a tibble with the same columns and data as murder_rates_death_penalty, but omits the column death_penalty_in_1971 and has only the rows for states that had the death penalty in 1971.

The first 2 rows of your table should look like this:

State	Year	Population	Murder Rate

Alabama	1960	3266740	12.4
Alabama	1961	3302000	12.9

```
murder_rates_death_penalty_active<- murder_rates_death_penalty %>%
  filter(death_penalty_in_1971 == TRUE) %>%
  select(State, Year, Population, `Murder Rate`)
murder_rates_death_penalty_active
```

```
## # A tibble: 1,936 × 4
##    State     Year Population `Murder Rate`
##    <chr>    <dbl>      <dbl>         <dbl>
##  1 Alabama  1960    3266740          12.4
##  2 Alabama  1961    3302000          12.9
##  3 Alabama  1962    3358000           9.40
##  4 Alabama  1963    3347000          10.2
##  5 Alabama  1964    3407000           9.30
##  6 Alabama  1965    3462000          11.4
##  7 Alabama  1966    3517000          10.9
##  8 Alabama  1967    3540000          11.7
##  9 Alabama  1968    3566000          11.8
## 10 Alabama  1969    3531000          13.7
## # … with 1,926 more rows
```

```
. = ottr::check("tests/part3_q3.R")
```

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

The null hypothesis doesn't specify *how* the murder rate changes; it only talks about increasing or decreasing. So, we will use the same test statistic you defined in Part II.

Question 3.4. Assign `observed_value_1972` to the value of the test statistic for the years 1971 to 1973 and the states in `murder_rates_death_penalty_active`.

Hint: You have already written nearly the same code in a previous part of this project.

```
observed_value_1972<- murder_rates_death_penalty_active %>%
  filter(Year == 1971 | Year == 1972 | Year == 1973) %>%
  group_by(State) %>%
  summarize(`Total differences` = sign(diff(`Murder Rate`,2))) %>%
  ungroup() %>%
  pull(`Total differences`) %>%
  test_stat()
observed_value_1972
```

```
## [1] 22
```

```
. = ottr::check("tests/part3_q4.R")
```

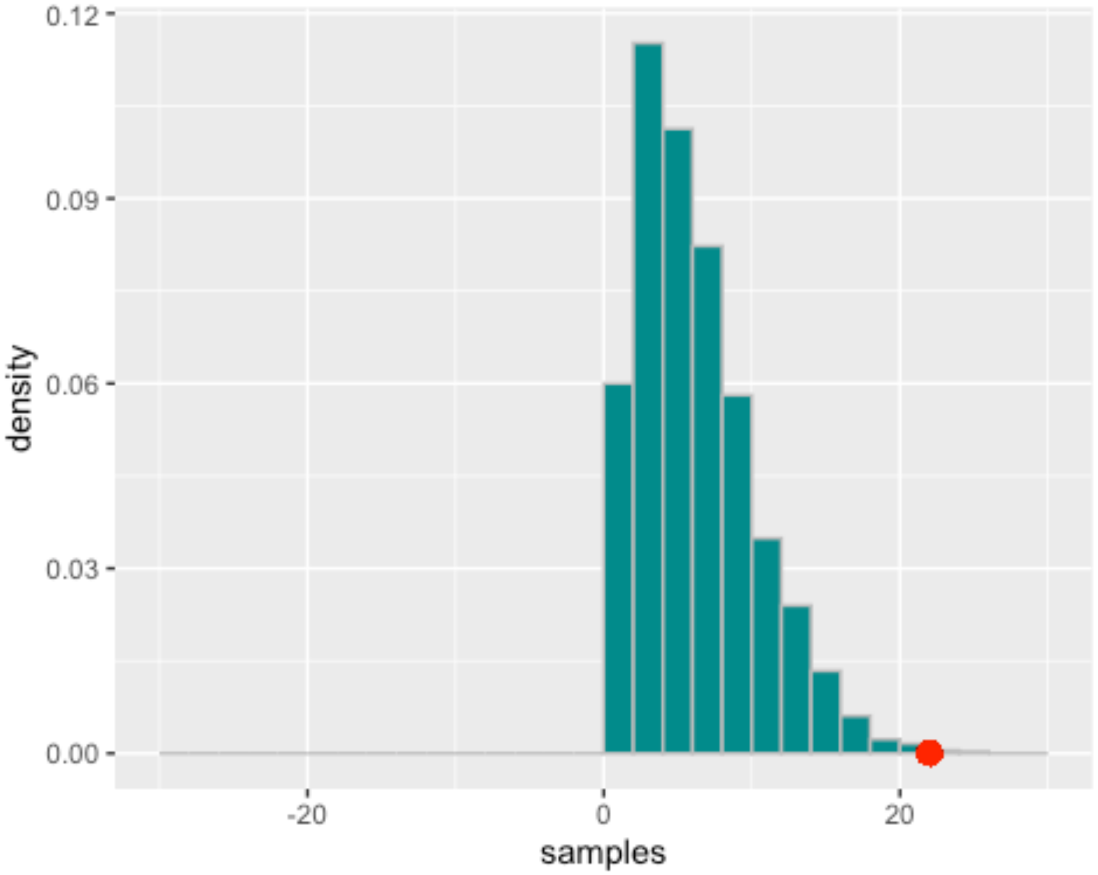
```
## All tests passed!
```

Question 3.5.: Simulate the test statistic under the null hypothesis 5,000 times. Assign the vector of results to the name `samples`.

```
samples<- replicate(n = 5000, simulate_under_null(44))

. = ottr::check("tests/part3_q5.R")

## All tests passed!
```

Question 3.6. Draw a sampling histogram of the statistic using `samples`. You may wish to annotate your plot with where the observed value is located on the histogram.

```
ggplot(tibble(samples)) +
  geom_histogram(aes(x = samples, y = ..density..), fill = "darkcyan",
                 color="gray", breaks=seq(-30, 30, 2), closed="left") +
  geom_point(aes(x = observed_value_1972, y = 0), size = 3, color = "red")
```



The following two questions complete the analysis:

Question 3.7. Compute a p-value using R code and assign the result to the name `death_penalty_p_value`.

```
death_penalty_p_value<- sum(samples >=
  observed_value_1972)/5000
death_penalty_p_value

## [1] 0.0016
```

Question 3.8. Draw a conclusion about the null and alternative hypotheses. Was the murder rate higher or lower during the two-year period when the death penalty was removed? Do we accept the null hypothesis? If not, what can we say?

Our p-value tells us that the chance of seeing a value at least as extreme as 22 states having murder rate increases from the years 1971-1973 is about 0.0016 or 0.16%. This is enough evidence to reject the null hypothesis at over a 99% significance level. We can conclude the alternative that the murder rate was more likely to increase as a result of the abolition of the death penalty.

Part IV: Further Evidence

So far, we have found evidence that when executions were outlawed (i.e. when death penalty was removed), the murder rate increased in many more states than we would expect from random chance. We have also seen that across all states and all recent years, the murder rate goes up about as much as it goes down over two-year periods.

These discoveries seem to support the claim that eliminating the death penalty increases the murder rate. Should we be convinced? Let's conduct some more tests to support our claim.

Conducting a test for this dataset required the following steps:

1. Select a tibble containing murder rates for certain states and all years,
2. Choose two years and compute the observed value of the test statistic,
3. Simulate the test statistic under the null hypothesis that increases and decreases are drawn uniformly at random, then
4. Compute a p-value by comparing the observed value to the sampling distribution.

This entire process can be expressed in a single function, called `run_test()`.

Question 4.1. Implement `run_test()`, which takes the following arguments:

- A tibble of murder `rates` for certain states, sorted by state and year like `murder_rates_death_penalty`, and
- the year when the analysis starts. The end year, or comparison group, is two years later.

The function prints out two values: (1) the "total increases minus the total decreases" and (2) the observed value of the test statistic (see Question 3.4). It returns the p-value for the observed value under the null hypothesis.

Hint: This part may appear intimidating at first but you can complete most of this question by copying code you wrote earlier.

```
run_test<- function(tibble,beginning_year){

  inc_minus_dec <- filter(tibble, Year == beginning_year | Year ==
beginning_year + 1 | Year == beginning_year + 2) %>%
  group_by(State) %>%
  summarize(`Total differences` = sign(diff(`Murder Rate`, 2))) %>%
  ungroup() %>%
  pull(`Total differences`) %>%
    sum()

  stat<- abs(inc_minus_dec)

  print(paste("The number of two-year increases minus decreases with sign: ",
inc_minus_dec))
  print(paste("Magnitude of change (our test stat): ", stat))

  p_value<- sum(samples >= stat)/length(samples)

  return(p_value)

}

# This example call should return what you found from Part III.
run_test(murder_rates_death_penalty_active, 1971)

## [1] "The number of two-year increases minus decreases with sign:  22"
## [1] "Magnitude of change (our test stat):  22"

## [1] 0.0016

. = ottr::check("tests/part4_q1.R")

## All tests passed!
```

The rest of the states

We found a dramatic increase in murder rates for those states affected by the 1972 Supreme Court ruling, but what about the rest of the states? There were six states that had already outlawed execution at the time of the ruling.

Question 4.2. Create a tibble called `murder_rates_death_penalty_inactive` with the same columns as `murder_rates_death_penalty_active` but only containing rows for the six states without the death penalty in 1971.

```
murder_rates_death_penalty_inactive<- murder_rates_death_penalty %>%
  filter(death_penalty_in_1971 == FALSE) %>%
  select(State, Year, Population, `Murder Rate`)
murder_rates_death_penalty_inactive

## # A tibble: 264 × 4
##    State   Year Population `Murder Rate`
##    <chr>  <dbl>      <dbl>         <dbl>
##  1 Alaska  1960     226167          10.2
##  2 Alaska  1961     234000          11.5
##  3 Alaska  1962     246000           4.5
##  4 Alaska  1963     248000          6.5
##  5 Alaska  1964     250000         10.4
##  6 Alaska  1965     253000          6.30
##  7 Alaska  1966     272000         12.9
##  8 Alaska  1967     272000          9.60
##  9 Alaska  1968     277000         10.5
## 10 Alaska  1969     282000         10.6
## # … with 254 more rows
```

```
. = ottr::check("tests/part4_q2.R")
```

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

Question 4.3. Perform the same test on this tibble (`murder_rates_death_penalty_inactive`). Assign the resulting p-value to the name `non_death_penalty_p_value`.

```
non_death_penalty_p_value <- run_test(murder_rates_death_penalty_inactive,
1971)
```

```
## [1] "The number of two-year increases minus decreases with sign:  1"
## [1] "Magnitude of change (our test stat):  1"
```

```
non_death_penalty_p_value
```

```
## [1] 0.88
```

Question 4.4. In one sentence, conclude whether their murder rates were also more likely to increase from 1971 to 1973.

Because the p-value for the non-death penalty states is very high, we do not have enough evidence to reject the null hypothesis and we cannot conclude that these states' murder rates were more likely to increase from 1971 to 1973.

The death penalty reinstated

In 1976, the Supreme Court repealed its ban on the death penalty in its rulings on a series of cases including Gregg v. Georgia, so the death penalty was reinstated where it was previously banned. This generated a second natural experiment. To the extent that the death penalty deters murder, reinstating it should decrease murder rates, just as banning it should increase them. Let's see what happened.

Question 4.5. Perform the test again on the tibble `murder_rates_death_penalty_active` but for the two-year period starting with 1975. Assign the resulting p-value to the name `death_penalty_reinstated_p_value`.

Hint: To check your result for the above test, figure out what your test statistic should be when there are 18 more decreases than increases, and verify that's the test statistic that was printed. Also, you should have found a p-value near 0.01. If your p-value is very different, go back and inspect your `run_test` implementation and your test statistic to make sure that it correctly produces low p-values when there are many more decreases than increases.

```
death_penalty_reinstated_p_value<-
run_test(murder_rates_death_penalty_active, 1975)

## [1] "The number of two-year increases minus decreases with sign:  -18"
## [1] "Magnitude of change (our test stat):  18"

death_penalty_reinstated_p_value

## [1] 0.0094
```

Question 4.6. Now we've analyzed states where the death penalty went away and came back, as well as states where the death penalty was outlawed all along. What do you conclude from the results of the tests we have conducted so far? Does all the evidence consistently point toward one conclusion, or is there a contradiction?

All the evidence we've collected is very convincing. By observing the states that never had a death penalty, we saw that murder rates increased about as often as they decreased. There were no significant changes in murder rates for these states across the years 1971-1977, as the p-value was too high (0.88). However, when we observed the states that originally had a death penalty and abolished it in the 1972 Supreme Court decision Furman vs Georgia, we saw that murder rates increased over the next two years with a p-value of 0.0016. Later, when the death penalty was reinstated in the 1976 Supreme Court case Gregg vs Georgia, murder rates decreased with a p-value of 0.01. Overall, the overwhelming evidence points to the conclusion that the existence of the death penalty lessens murder incentives in the US.

Part V: Visualization.

While our analysis appears to support the conclusion that the death penalty deters murder, a 2006 Stanford Law Review paper argues the opposite: that historical murder rates do **not** provide evidence that the death penalty deters murderers.

To understand their argument, we turn to visualization. In fact, we've done this whole analysis backward – we should always visualize first and then ask precise rigorous questions later!

What plot should we draw?

We know that we want to compare murder rates of states with and without the death penalty. We know we should focus on the period around the two natural experiments of 1972 and 1976, and we want to understand the evolution of murder rates over time for those groups of states. It might be useful to look at other time periods, so let's plot them all for good measure.

Question 5.1. Create a tibble from `murder_rates_death_penalty` that gives the average murder rate per year with respect to states that had the death penalty in 1971 and those that did not. Call this tibble `average_murder_rates`. It should have 3 columns:

- Year, the year
- Group, indicating if the figure reported for the row is for "Death penalty states" or "No death penalty states".
- avg_murder_rate, the average murder rate for that group in that year.

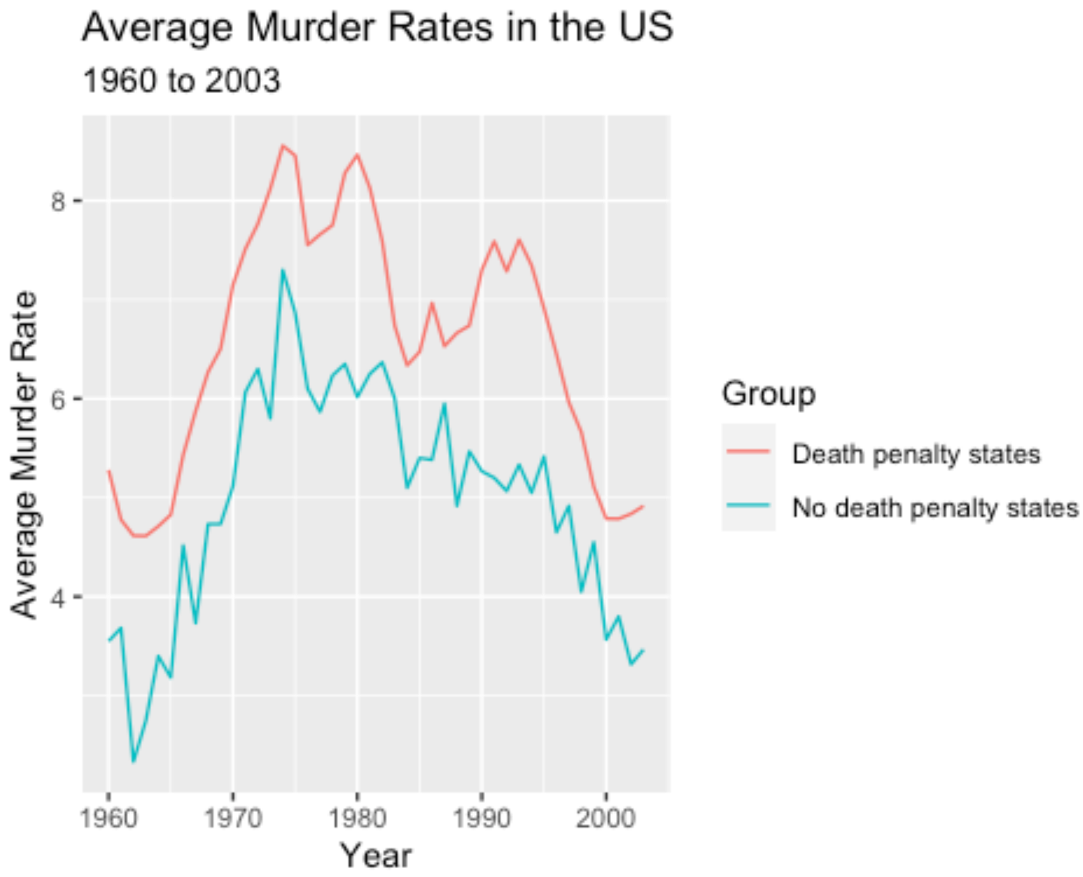
Its first few rows should look like:

Year	Group	avg_murder_rate
1960	No death penalty states	3.55
1960	Death penalty states	5.279
1961	No death penalty states	3.683
1961	Death penalty states	4.77

```
average_murder_rates<- murder_rates_death_penalty %>%
  mutate(Group = replace(death_penalty_in_1971, death_penalty_in_1971 ==
TRUE, "Death penalty states")) %>%
  mutate(Group = replace(Group, Group == FALSE, "No death penalty states"))
%>%
  group_by(Year, Group) %>%
  summarize(avg_murder_rate = mean(x = `Murder Rate`)) %>%
  arrange(desc(Group), .by_group=TRUE) %>%
  ungroup()

average_murder_rates

## # A tibble: 88 × 3
##     Year Group                   avg_murder_rate
##    <dbl> <chr>                             <dbl>
##  1  1960 No death penalty states            3.55
##  2  1960 Death penalty states               5.28
##  3  1961 No death penalty states            3.68
##  4  1961 Death penalty states               4.78
##  5  1962 No death penalty states            2.33
##  6  1962 Death penalty states               4.62
##  7  1963 No death penalty states            2.75
##  8  1963 Death penalty states               4.61
##  9  1964 No death penalty states            3.40
## 10  1964 Death penalty states               4.71
## # … with 78 more rows

   # source for replace() in mutate(): https://statisticsglobe.com/r-replace-
value-of-data-frame-variable-dplyr-package
   # source for .by_group in arrange():
https://stackoverflow.com/questions/43832434/arrange-within-a-group-with-
dplyr

. = ottr::check("tests/part5_q1.R")

## All tests passed!
```

Question 5.2. Generate a line plot using `average_murder_rates`. There should be two lines shown in a single plot: one showing the trend of average murder rates in death penalty states and another the trend of average murder rates in non-death penalty states. As for all visualizations, text should be legible, axes labeled appropriately, have a title, etc.

```
ggplot(average_murder_rates, aes(x=Year, y=`avg_murder_rate`, color=Group)) +
  geom_line()+
  labs(title = "Average Murder Rates in the US",
       subtitle = "1960 to 2003",
       y= "Average Murder Rate")
```



Let's bring in another source of information: Canada. Run the following cell to bring this data into the environment:

```
canada_rates <- read_csv("data/canada.csv") %>%
  select(Year, Homicide) %>%
  mutate(Group = 'Canada') %>%
  rename("avg_murder_rate" = "Homicide")
canada_rates

## # A tibble: 53 × 3
##      Year avg_murder_rate Group
##     <dbl>           <dbl> <chr>
##   1  1963            1.32 Canada
##  2  1964          1.31 Canada
##  3  1965          1.41 Canada
##  4  1966          1.25 Canada
##  5  1967          1.66 Canada
##  6  1968          1.81 Canada
##  7  1969          1.86 Canada
##  8  1970          2.19 Canada
##  9  1971          2.15 Canada
## 10  1972          2.34 Canada
## # … with 43 more rows
```

Question 5.3. Append the average murder rate figures for Canada to the data that we have in `average_murder_rates` (note that the columns in both tibbles are the same). Put another way, we would like to "stack" one tibble on top of the other so that we have one larger tibble with figures for both countries. Call this resulting tibble `average_murder_rates_with_canada`.

Hint: If you are unsure how to do this, review Section 4.4.2 for hints on pulling data from multiple sources.

```
average_murder_rates_with_canada<- bind_rows(average_murder_rates,
canada_rates)
average_murder_rates_with_canada
```

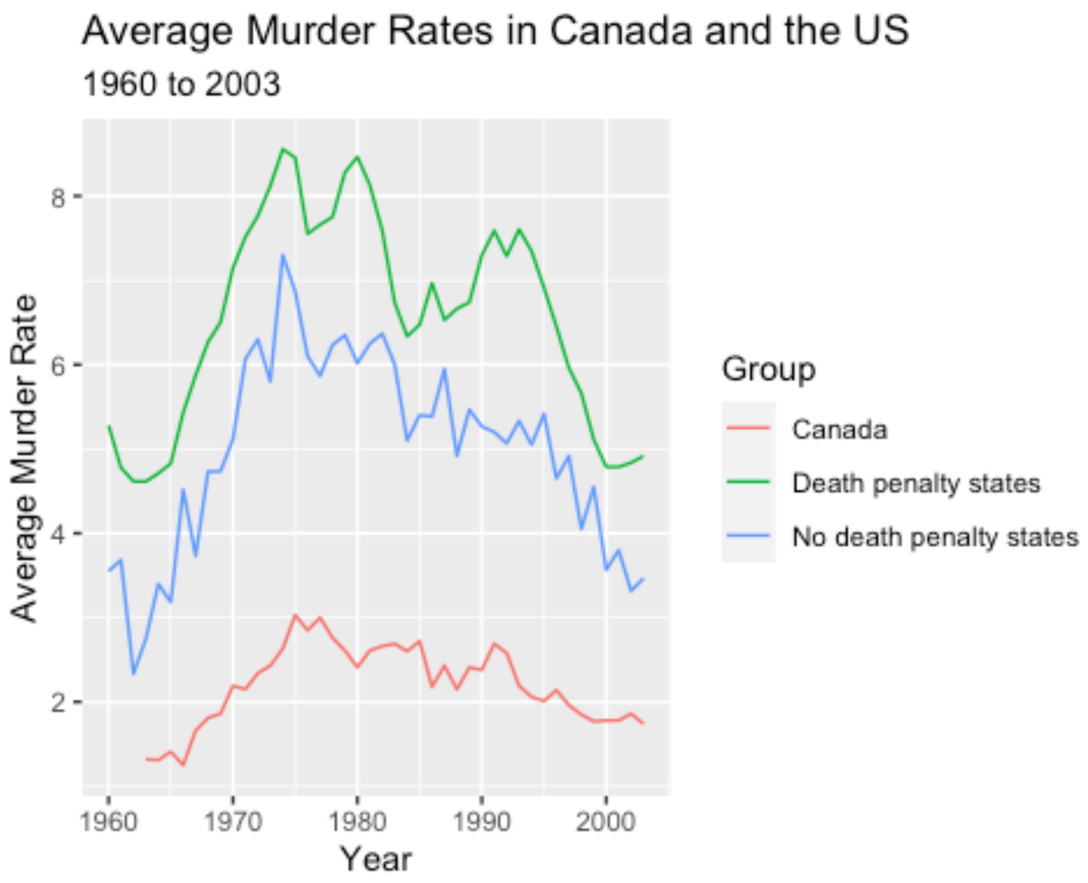
```
## # A tibble: 141 × 3
##     Year Group                   avg_murder_rate
##    <dbl> <chr>                             <dbl>
##  1  1960 No death penalty states            3.55
##  2  1960 Death penalty states               5.28
##  3  1961 No death penalty states            3.68
##  4  1961 Death penalty states               4.78
##  5  1962 No death penalty states            2.33
##  6  1962 Death penalty states               4.62
##  7  1963 No death penalty states            2.75
##  8  1963 Death penalty states               4.61
##  9  1964 No death penalty states            3.40
## 10  1964 Death penalty states               4.71
## # … with 131 more rows
```

```
. = ottr::check("tests/part5_q3.R")
```

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

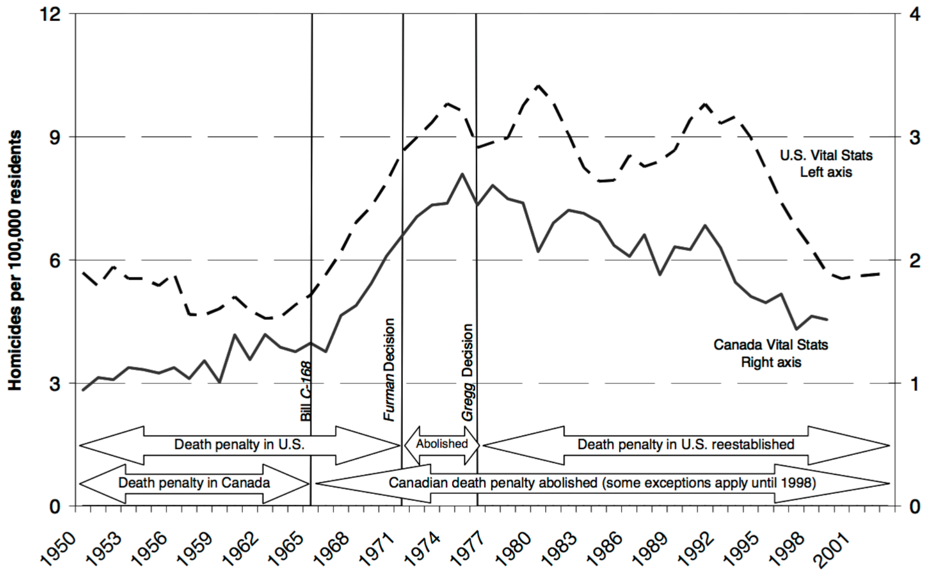
Question 5.4. Re-create your line plot, but now including the Canada figures. As for all visualizations, all text should be legible, axes labeled appropriately, have a title, etc.

```
average_murder_rates_with_canada_up_to_2003 <-
average_murder_rates_with_canada %>% slice_head(n=129)
ggplot(average_murder_rates_with_canada_up_to_2003, aes(x=Year,
y=`avg_murder_rate`, color=Group)) +
  geom_line()+
  labs(title = "Average Murder Rates in Canada and the US",
       subtitle = "1960 to 2003",
       y= "Average Murder Rate")
```



This plot corresponds closely to a visualization from the paper.

Figure 2. Homicide Rates and the Death Penalty in the United States and Canada



Canada has not executed a criminal since 1962. Since 1967, the only crime that can be punished by execution in Canada is the murder of on-duty law enforcement personnel. The paper states, "The most striking finding is that the homicide rate in Canada has moved in virtual lockstep with the rate in the United States."

Question 5.5. How does this visualization indicate that the death penalty is not an important factor in determining the murder rate?

This visualization shows that the trends for murder rates are nearly the same for all three groups. The trend shows a steady increase (on average) in murder rates from 1960-1977 across death penalty states, non death penalty states, and even Canada. Then it shows an average steady decrease from 1977 to 2003 for all three groups as well. Because these trends are very similar across groups in the visualization, this indicates that the death penalty is not an important factor in determining the murder rate.

Question 5.6. What assumption(s) did we make in this project that led us to conclude that the death penalty deterred murder, when the above visualization tells a different story?

In order to compute our p-values, we assumed that the null hypothesis was true. That is, we assumed that whether the murder rate increased or decreased in each state was like the flip of a fair coin. However, we can see from the visualization that there were steady trends of increasing and decreasing periods across groups. We did not take this into consideration when originally stating our null hypothesis.

You finished the project – well done!!

Make sure that all automatic tests given are passing and that you are fully confident in the answers you have given. As with all your assignments, submit both the final .Rmd notebook file *AND* a generated PDF document before the final deadline noted at the top of this notebook and on Gradescope.