assignment_04

2023-09-17

Step by Step

The correct order is as follows: 5, 1, 3, 2, 4, 6.

So, rewriting the steps in order would look like this:

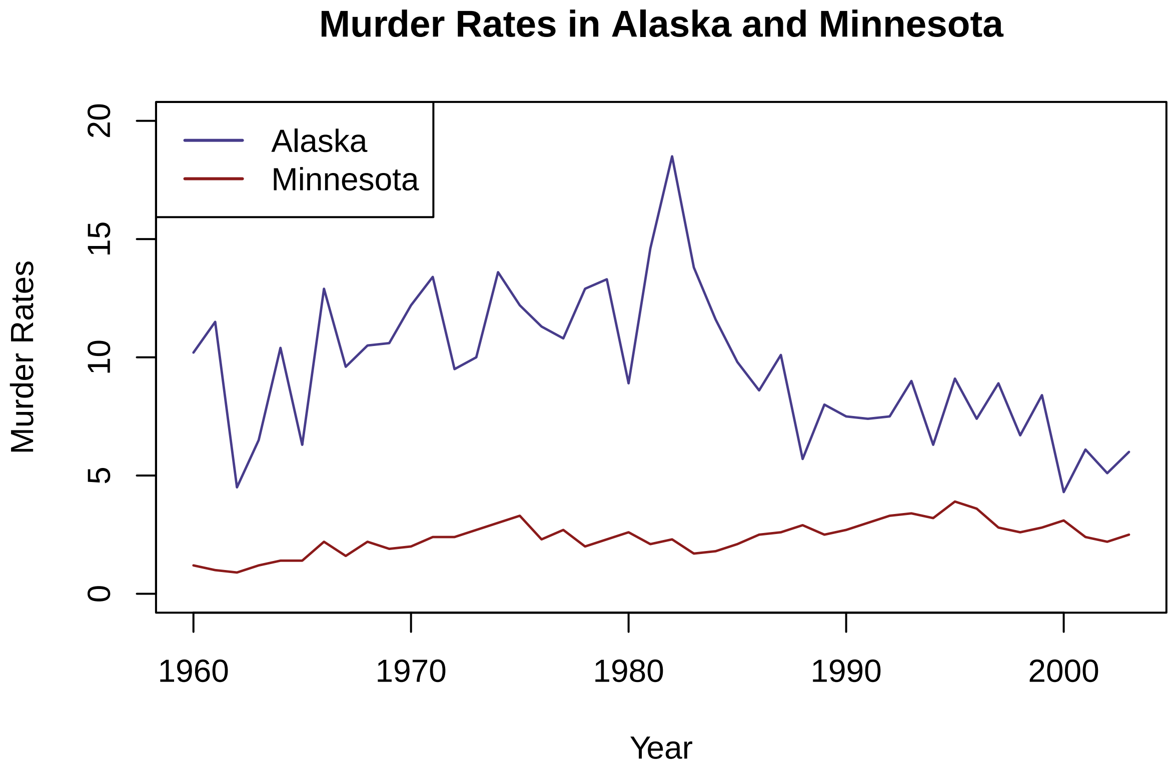
1. Define a null and alternate model
2. Choose a test statistic (typically the difference in means between two categories)
3. Find the value of the observed test statistic
4. Shuffle the labels of the original sample, find your simulated test statistic, and repeat many times
5. Calculate the p-value based off your observed and simulated test statistics
6. Use the p-value and p-value cutoff to draw a conclusion about the null hypothesis

How to Get Away with Murder

Reading the .csv file and creating the required dataframe:

```
crime_rates_raw<- read.csv("./data/crime_rates.csv")
crime_rates_murder <- crime_rates_raw[c("State", "Year", "Population", "Murder.Rate")]

#separating Alaska and Minnesota to make the plot code neater later
#I can probably just have included this subsetting within the plot but messy messy nvm
murder_rates_alaska <- crime_rates_murder[crime_rates_murder$State == "Alaska",]
murder_rates_minnesota <- crime_rates_murder[crime_rates_murder$State == "Minnesota",]
```

Exploring the trends in Alaska and Minnesota:



Yikes, why does Alaska have such a high murder rate.

Death Penalty Goes to Die

Yes, we should use A/B testing here. There wouldn't be a point doing this homework if we aren't going to use A/B testing right LOL. Jokes aside, it does make sense, since we want to compare the murder rates before and after the death penalty was abolished to see if any increase in murder is just by chance or actually due to the death penalty abolishment. We choose "A" to be the states with death penalty post ban, and "B" to be the states with death penalty pre ban.

Death Penalty

Here, we merge the rates with the death penalty information and extract a data frame that consists of data of states with the death penalty before the abolition:

```
death_penalty <- read.csv("./data/death_penalty.csv")
crime_rates_all_with_death <- merge(x = crime_rates_murder,
                                    y = death_penalty,
                                    by = "State")

preban_rates <- crime_rates_all_with_death[crime_rates_all_with_death$Death.Penalty == TRUE
                                           & crime_rates_all_with_death$Year == 1971, ]
```

Viewing the first few rows as requested:

```
##            State Year Population Murder.Rate Death.Penalty
## 12       Alabama 1971    3479000        15.1          TRUE
## 100      Arizona 1971    1849000         6.7          TRUE
## 144     Arkansas 1971    1944000        10.5          TRUE
## 188   California 1971   20223000         8.1          TRUE
## 232     Colorado 1971    2283000         6.5          TRUE
## 276 Connecticut 1971    3081000         3.1          TRUE
```

Now, creating a data frame for the post ban rates:

```
postban_rates <- crime_rates_all_with_death[crime_rates_all_with_death$Death.Penalty == TRUE
                                           & crime_rates_all_with_death$Year == 1973, ]
#Ok, couldn't figure out how to do this so I just brute forced it oops
Death.Penalty <- logical(length = nrow(postban_rates))
postban_rates <- cbind(postban_rates[,-5], Death.Penalty)
```

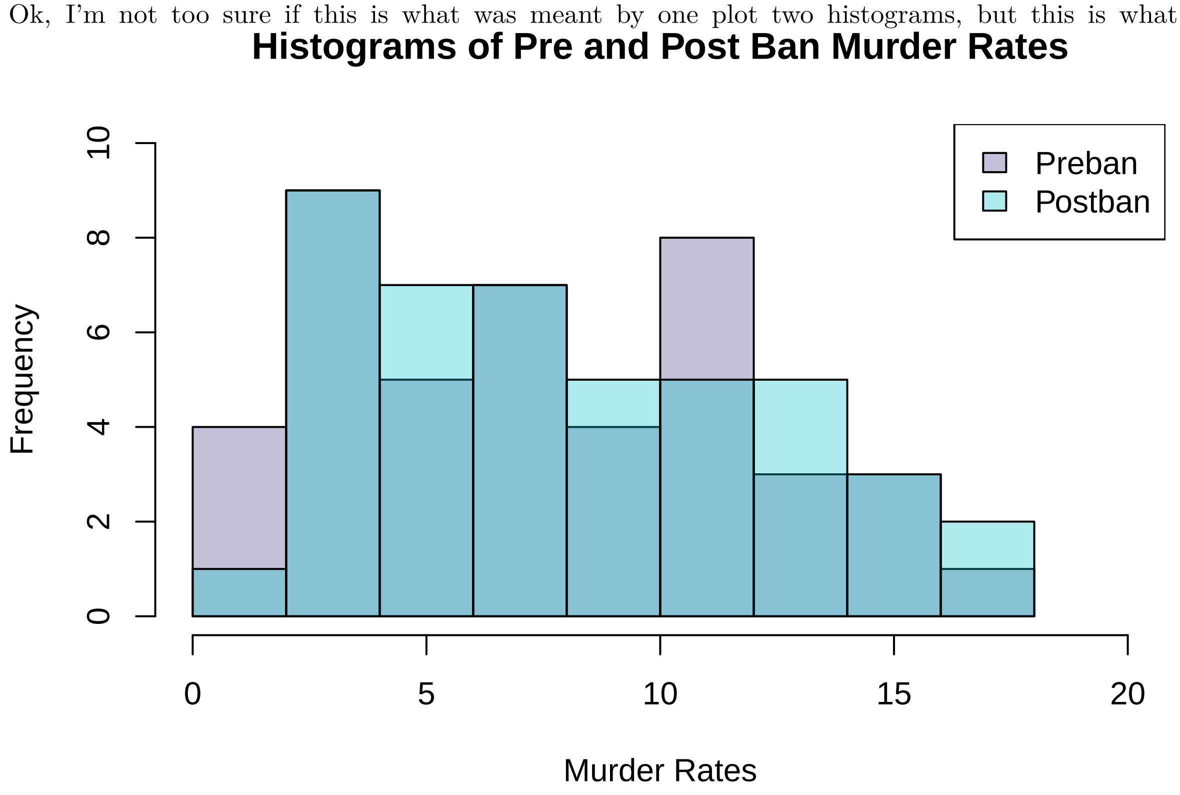
Viewing the postban rates:

```
##           State Year Population Murder.Rate Death.Penalty
## 14       Alabama 1973    3539000        13.2         FALSE
## 102      Arizona 1973    2058000         8.1         FALSE
## 146     Arkansas 1973    2037000         8.8         FALSE
## 190   California 1973   20601000         9.0         FALSE
## 234     Colorado 1973    2437000         7.9         FALSE
## 278  Connecticut 1973    3076000         3.3         FALSE
```

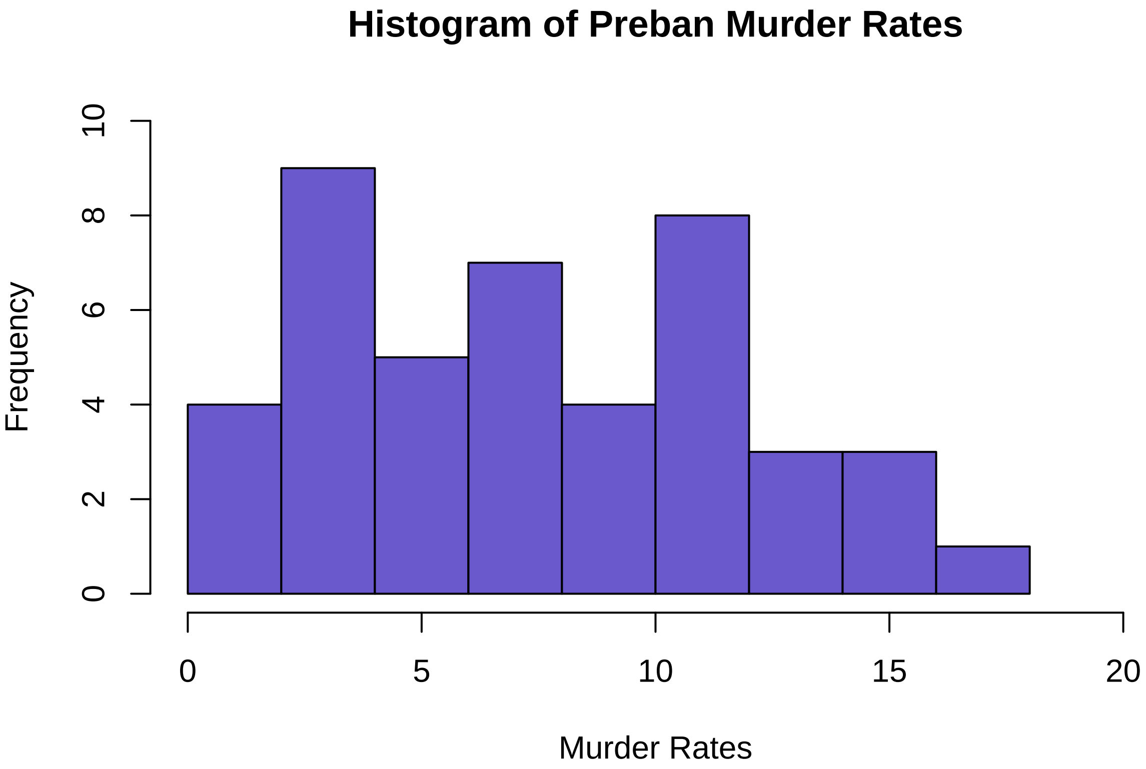
Combining the two dataframes:

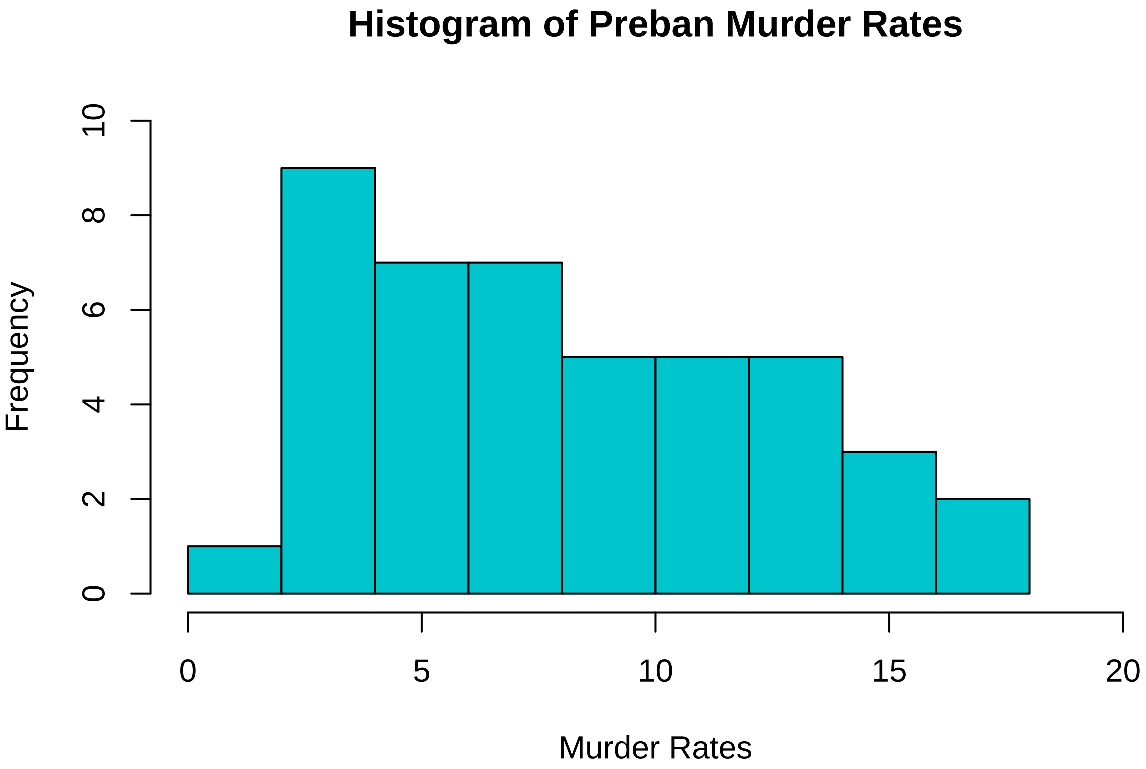
```
#I assume it's just combining them vertically since I can't fathom any other way to do it
change_in_death_rates <- rbind(preban_rates, postban_rates)
```

Ok, I'm not too sure if this is what was meant by one plot two histograms, but this is what I did:



I have learned that it is TWO SEPARATE HISTOGRAMS so here's the plots for those, but I'm leaving the one above in here too because I already spent 30 minutes on it:





Here, we calculate the murder rate averages and put them in the dataframe as required:

```
average_with<- mean(preban_rates$Murder.Rate)
average_without <- mean(postban_rates$Murder.Rate)
```

```
Murder.Rate.Averages <- c(average_without, average_with)
Death.Penalty <- c(FALSE, TRUE)

#I am lazy to rename the columns
rate_means <- data.frame(Death.Penalty, Murder.Rate.Averages)
```

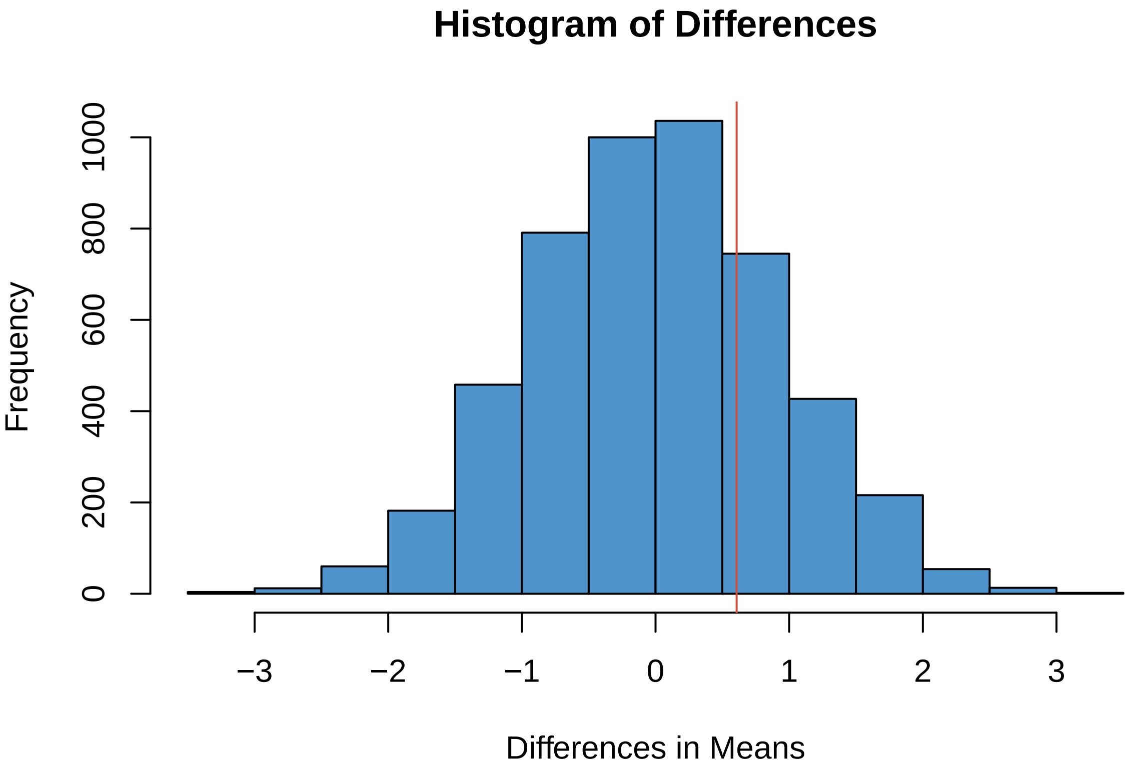
I think there's something incorrect in the phrasing of the question. The question asked for "the average murder rates for the states that had the death penalty and the states that didn't have the death penalty", which seems to imply one set of states for one average, and another set of states for the other average. Furthermore, I couldn't get the correct values if I included any state that did not have a death penalty. The values that I've found are the average murder rates for states that had the penalty before and after abolishment of the death penalty. I'm not sure if this is exactly what the question wanted, but it matches the table in the assignment sheet.

A useful test statistic would thus be the difference in the average murder rates before and after abolishment. Smaller differences in the test statistic would help us reject the null hypothesis, that abolishment did nothing, and accept the alternate hypothesis, that it did made murder rates go up.

Here, we are setting up to shuffle the labels, and then performing the shuffling 5000 times and calculating difference in the means as requested:

```
observed_difference <- rate_means[1,2] - rate_means[2,2]

#data = dataframe, label = murder rates, group_label = death penalty
#The $ way of subsetting won't work, because we are passing in a string
#data["column name"] to subset with column names
calculate_difference <- function(data, value_label, group_label) {
  average_death <- mean(data[data[group_label] == TRUE, value_label])
  average_no_death <-mean(data[data[group_label] == FALSE, value_label])
  return(average_death - average_no_death)
}

nrows <- nrow(change_in_death_rates)
simulate_and_test_statistic <- function() {
  shuffled <- change_in_death_rates
  shuffled_rows <- sample(nrows)
  shuffled$Death.Penalty <- (change_in_death_rates$Death.Penalty)[shuffled_rows]
  return(calculate_difference(shuffled, "Murder.Rate", "Death.Penalty"))
}

wrapper <- function(x){
  simulate_and_test_statistic()
}

differences <- sapply(1:5000, wrapper)
```

Plotting the test statistics in a histogram:



Now, we calculate the p-value:

```
## [1] 0.2556
```

Using a p-value cutoff of 5%, we cannot reject the null hypothesis, as our p-value is way bigger than 5%. Basically, we cannot conclude that the death penalty abolishment increased the rates of murder as the p-value means there is around a 26% chance (for the simulation I ran) of getting as extreme or more extreme result as the observed difference. It means that the murder rates after abolishing the death penalty did not really change, so the death penalty was likely not a good deterrent for murder since people were murdering at similar rates when the death penalty was in action and when it was abolished. However, it is possible that, because of randomness, the effects of the death penalty abolishment cannot be seen from the two years that we chose to analyse. Perhaps it just happened that 1971 had exceptionally high murder rates. We could examine other years preban to check.