Course Project - Part II: Comparing Samples with Resampling Techniques

Ken Wood

7/25/2024

In Part Two of the course project, you will first use resampling methods to assess the uncertainty around the social mobility scores of highly selective private and public colleges. Then, you'll compare social mobility scores of colleges in two groups: highly selective public colleges and highly selective private colleges.

To load this data set and view a boxplot of the social mobility score (mobility_rate) based on whether the school is public or private (tier_name), run the following code chunk:

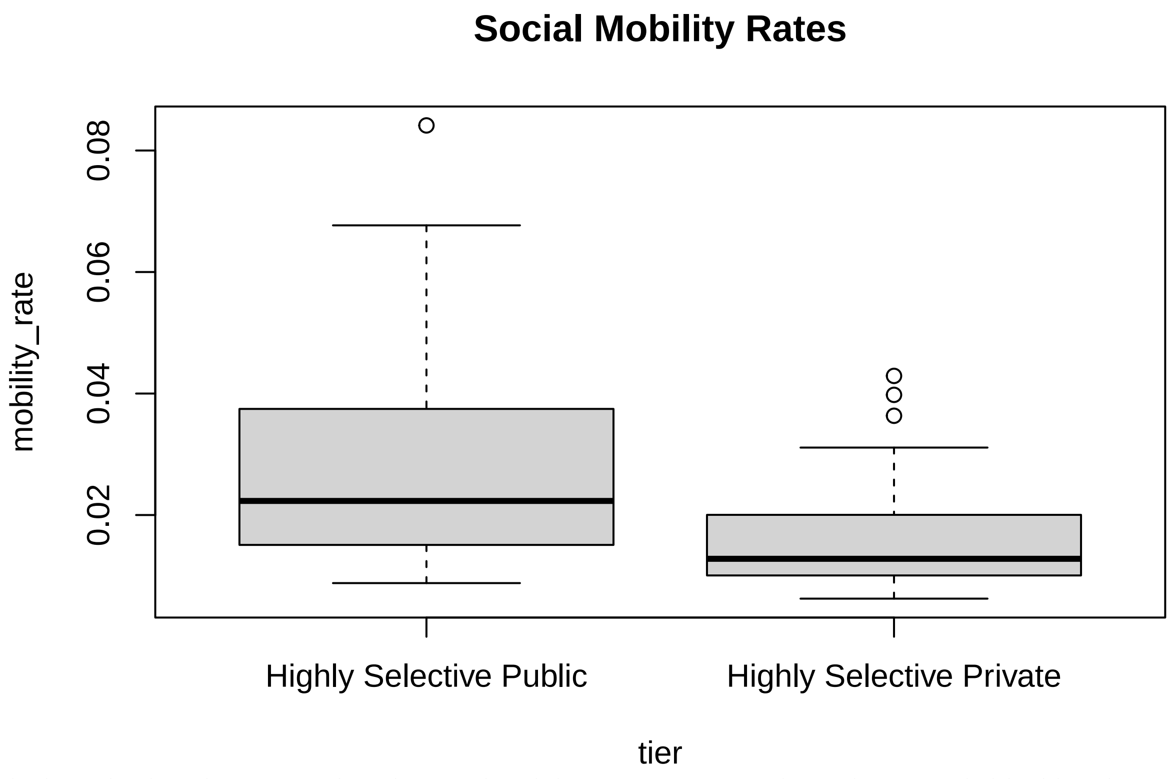
```r
# eCornell Hex Codes:
crimson = '#b31b1b'   # crimson
lightGray = '#cecece' # lightGray
darkGray = '#606366'  # darkGray
skyBlue = '#92b2c4'   # skyblue
gold = '#fbb040'      # gold
ecBlack = '#393f47'   # ecBlack

school = read.csv('mrc_table2.csv', header = TRUE, check.names = FALSE)
school = school[,names(school) %in%
                  c('name', 'type', 'tier', 'tier_name', 'mr_kq5_pq1',
                    'par_median', 'k_median')]
names(school)[5:7] <- c("mobility_rate", "parent_income", "student_income")

school <- school[c(which(school$tier == 4), which(school$tier == 3)),]

HS_private_scores <- school$mobility_rate[school$tier_name == 'Highly selective private']
HS_public_scores <- school$mobility_rate[school$tier_name == 'Highly selective public']

boxplot(mobility_rate ~ tier, data = school, main = 'Social Mobility Rates', names = c('Highly Selectiv
```



As you look at this boxplot, notice that the social mobility rate among students that attended highly selective public schools is higher than that of students who attended highly selective private schools.

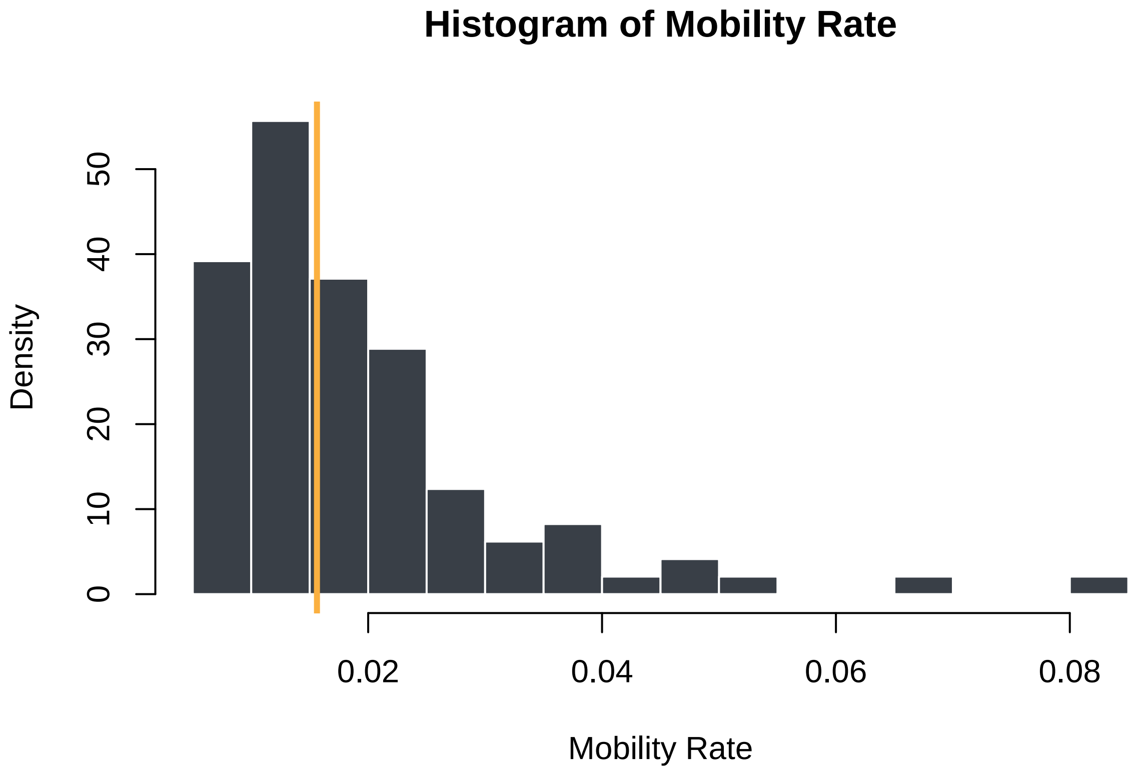
However, this boxplot is based on a small subset of all colleges in the US. Use the techniques you practiced in this module to assess the uncertainty of the results you see in this boxplot. Assessing the uncertainty around these results will help you determine whether the outcome you see in the boxplot above can be generalized to all colleges in the United States.

Step 1

Draw a histogram of social mobility scores of all colleges in this data set. Describe the shape of this histogram. Calculate the median social mobility score of this data set.

```r
hist(school$mobility_rate, breaks = 20, freq = FALSE, col = ecBlack, border = 'white',
xlab = 'Mobility Rate',
main = 'Histogram of Mobility Rate')

abline(v = median(school$mobility_rate), col = gold, lwd = 3)
```



```
median(school$mobility_rate)
```

```
## [1] 0.01562073
```

Step 2

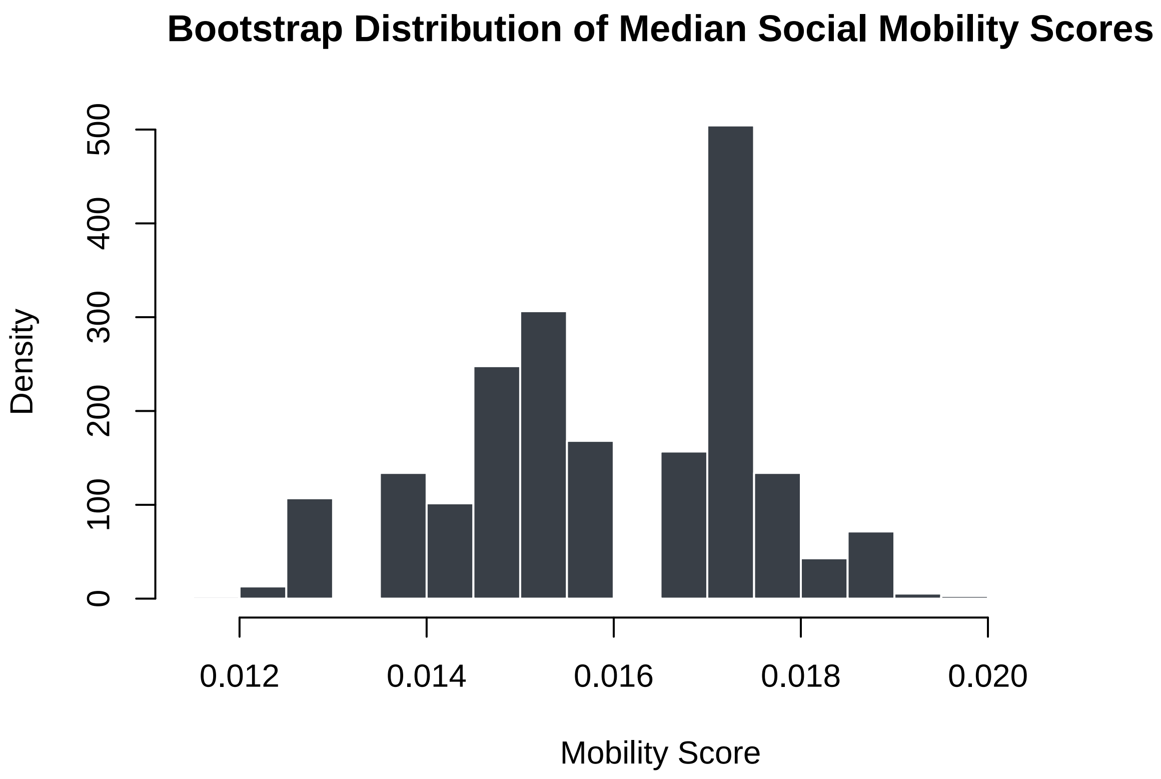
Complete the code chunk below to draw 10,000 bootstrap samples from this data set and create a bootstrap distribution of median social mobility scores. Complete the lines of code that say # COMPLETE after them.

```
set.seed(1)
B = 10000

store_median = rep(0, B)

for (n in 1:B){
  boot.id = sample(1:97, size = 97, replace = TRUE)
  school.boot = school[boot.id,]
  store_median[n] = median(school.boot$mobility_rate)
}

hist(store_median, breaks = 20, freq = FALSE, col = ecBlack,
    border ='white', main = 'Bootstrap Distribution of Median Social Mobility Scores',
    xlab = 'Mobility Score')
```



##
Step 3 Use the bootstrap distribution of median social mobility score you created in Step 2 to calculate a 95% confidence interval around the median social mobility scores.

```
ci.95 = quantile(store_median, probs = c(0.025, 0.975))
ci.95
```

```
##       2.5%      97.5%
## 0.01280050 0.01857753
```

Step 4

You want to compare the social mobility scores of highly selective public schools (HS_public_scores) with highly selective private schools(HS_private_scores) to see if they are truly different. What sample statistic would you build your distribution around? Calculate the observed sample statistic and store the value as obs_stat.

```
obs_stat <- mean(HS_public_scores) - mean(HS_private_scores)
```

Step 5

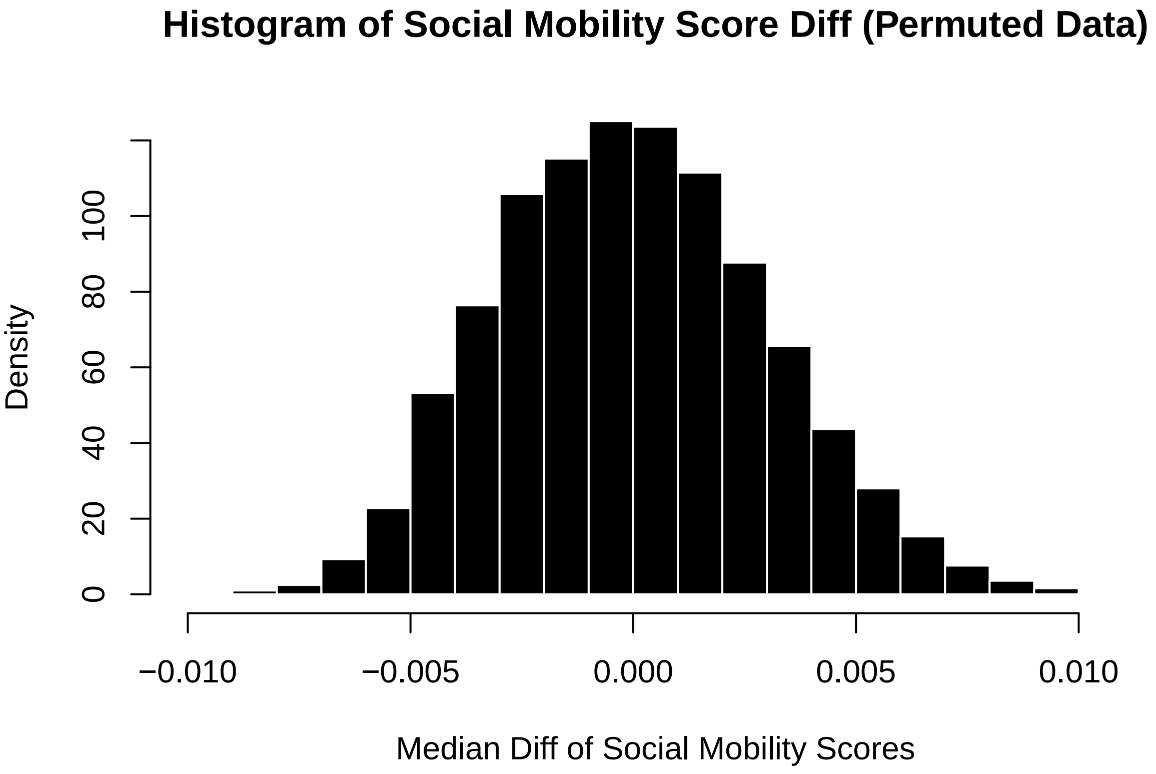
Complete the code chunk below to generate 10,000 permutated data sets and compare the social mobility scores of highly selective public colleges with those of highly selective private schools in each of them, then create a distribution of the sample statistics. Complete the lines of code that say # COMPLETE after them.

```
set.seed(1)
P = 10000
store_median_diff = rep(0, P)

for (n in 1:P){
  school.perm = school
  school.perm$mobility_rate = sample(school$mobility_rate, replace = FALSE)
  school.perm.prv = school.perm$mobility_rate[school.perm$tier_name == 'Highly selective private']
  school.perm.pub = school.perm$mobility_rate[school.perm$tier_name == 'Highly selective public']

  store_median_diff[n] = mean(school.perm.pub, na.rm=TRUE) - mean(school.perm.prv, na.rm=TRUE)
}

hist(store_median_diff, breaks = 20, freq = FALSE, col = 'black', border = 'white',
     xlab = 'Median Diff of Social Mobility Scores',
     main = 'Histogram of Social Mobility Score Diff (Permuted Data)')
```

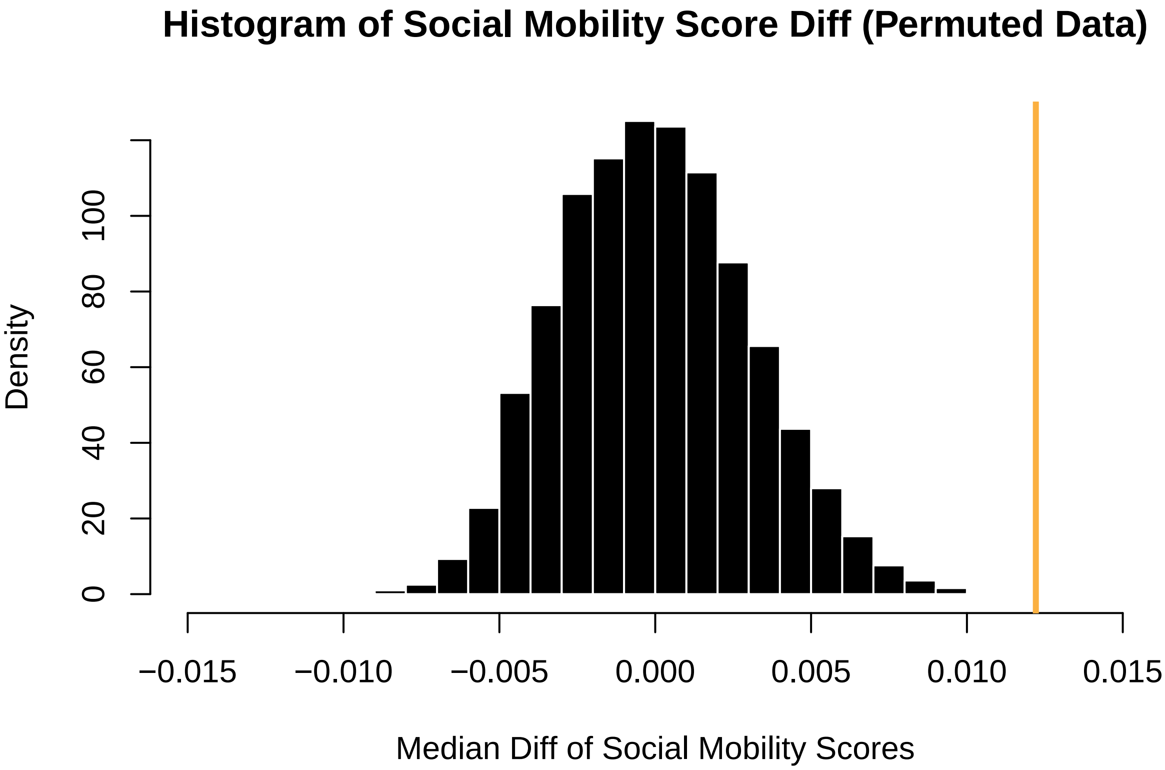


Step 6

Plot the observed sample statistic on the histogram, then calculate the p-value of the observed statistic.

```
hist(store_median_diff, breaks = 20, xlim = c(-0.015,0.015), freq = FALSE, col = 'black', border = 'whi
     xlab = 'Median Diff of Social Mobility Scores',
     main = 'Histogram of Social Mobility Score Diff (Permuted Data)')

abline(v = obs_stat, col = gold, lwd = 3)
```



```
# Calculate p-value
mean(abs(store_median_diff) >= abs(obs_stat)) # =0
```

```
## [1] 0
```

Step 7

Based on your results in Step 6, determine whether you think social mobility scores truly differ between highly selective public and private colleges in the population. Briefly explain your answer.

Since the p-value is < 0.05, we must reject the null hypothesis that there is no statistically significant difference in the means of the median mobility scores of public institutions vs. the mean of median mobility scores of private institutions. Therefore, we can conclude that social mobility scores DO differ between highly selective public and private colleges in the population.