Bootstrap Sampling: Exposure to Inequality (replication)

Brianna Hayes

2023-12-17

> Sands, Melissa L. 2017. "Exposure to inequality affects redistribution." *Proceedings of the National Academy of Sciences*, 114(4): 663-668.

In this experiment, actors were either white or black, were dressed either to embody affluence or poverty. The actors were stationed at street corners in Boston and standing within 20 feet of a petitioner. The petitioner would ask every third person walking by if they would like to sign a petition, about the millionaire tax (the imposition of an additional 4% tax on individuals with an income greater than $1 million). The experiment aimed to see if the race of the nearby actor, or how they were dressed, could possibly influence the decision to sign the millionaire tax petition.

```r
setwd('/Users/briannahayes/Documents/Jobs/R')
exposure <- read.csv('exposure_replication_data.csv', stringsAsFactors = TRUE)
exp <- subset(exposure, exposure$Mill_tax==1)

exp$costume[exp$pooractor == 1] <- 'Poor'
exp$costume[exp$pooractor != 1] <- 'Affluent'

exp$race_actor[exp$blackactor == 1] <- 'Black'
exp$race_actor[exp$blackactor != 1] <- 'White'
```

1. **Variables**

Independent variable 1: race_actor, race of the nearby actor (Black, White)

Independent variable 2: costume, how the actor was dressed (Affluent, Poor)

Dependent variable: Mill_tax_signed, Decision to sign the millionaire tax petition (1=signed, 0=abstained)

2. **Difference in means between black and white actors**

```r
mean_poor <- mean(exp$Mill_tax_signed[exp$costume=='Poor'])
mean_aff <- mean(exp$Mill_tax_signed[exp$costume=='Affluent'])
mean_poor
```

```
## [1] 0.06646972
```

```r
mean_aff
```

```
## [1] 0.08662614
```

```r
pop_dim <- mean_aff - mean_poor
pop_dim
```

```
## [1] 0.02015642
```

The mean proportion of those who were petitioned about the millionaire's tax, that signed the millionaire's tex petition upon seeing a poor actor, is .066. The proportion of those who signed it after seeing an affluent actor is .086. Our target statistic here is the difference between these means, which is .02. Thus, when people were petitioned about the millionaire's tax, it doesn't appear as though the decision to sign it was highly dependent upon whether they saw a poor or affluent actor.

3. **Generating 1000 bootstrap replications of the estimated difference in means and plotting them**

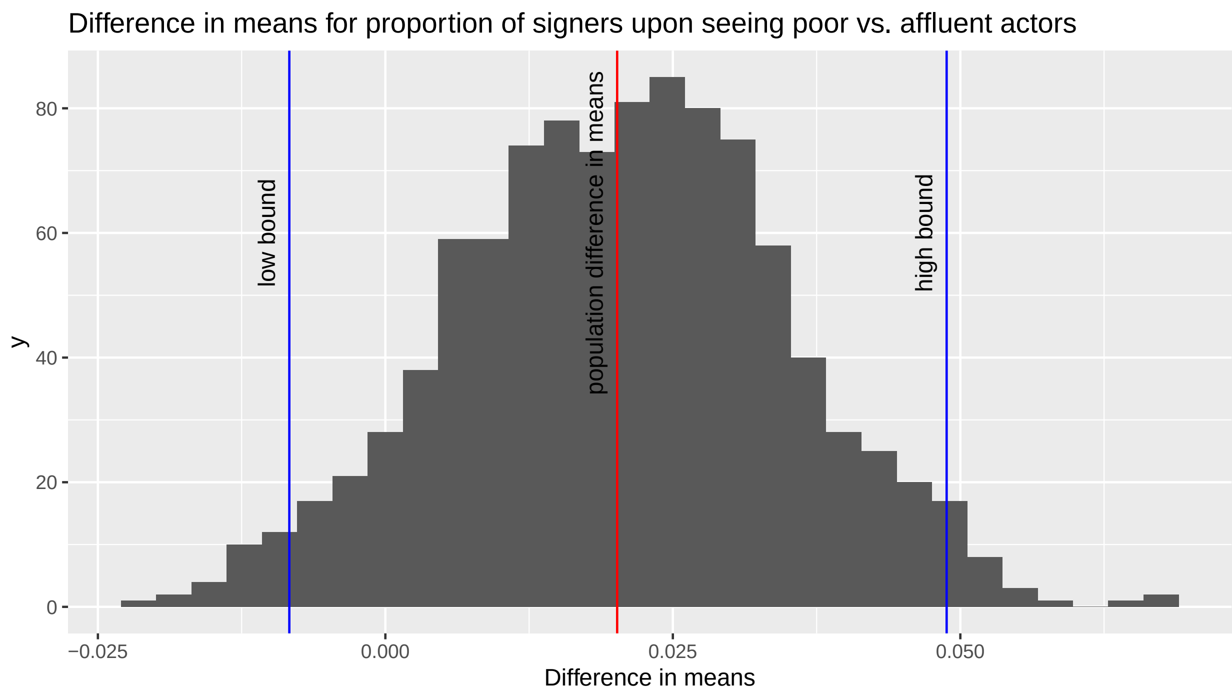
```r
## Bootstrap replications
bootstrap <- c()
for (i in 1:1000) {
  row_nums <- sample(1:nrow(exp), nrow(exp), replace=T)
  samp_exp <- exp[row_nums,]
  samp_mean_poor <- mean(samp_exp$Mill_tax_signed[samp_exp$costume=='Poor'])
  samp_mean_aff <- mean(samp_exp$Mill_tax_signed[samp_exp$costume=='Affluent'])
  bootstrap <- append(bootstrap, samp_mean_aff - samp_mean_poor)
}

## Confidence interval bounds
low_bound <- quantile(bootstrap, .025)
high_bound <- quantile(bootstrap, .975)

low_bound
```

```
##         2.5%
## -0.008352771
```

```r
high_bound
```

```
##      97.5%
## 0.04881088
```

```r
## Bootstrap distribution
ggplot() +
  geom_histogram(aes(bootstrap)) +
  geom_vline(xintercept=pop_dim, col='red') +
  geom_vline(xintercept=low_bound, col='blue') +
  geom_vline(xintercept=high_bound, col='blue') +
  ggtitle('Difference in means for proportion of signers upon seeing poor vs. affluent actors') +
  xlab('Difference in means') +
  annotate('text', x=low_bound - .002, y=60, label='low bound', angle=90) +
  annotate('text', x=high_bound - .002, y=60, label='high bound', angle=90) +
  annotate('text', x=pop_dim - .002, y=60, label = 'population difference in means', angle=90)
```



The low bound at the 2.5 percentile is -.007. The high bound of the confidence interval at the 97.5 percentile is .048. Thus, this reference distribution contains 0 in the confidence interval. When the confidence interval contains 0, there is very possible that there will be no difference between means in a given simulation.

3. **Difference in difference in means and regression**

```
## White actor difference in means by costume
wa_p <- mean(exp$Mill_tax_signed[exp$race_actor=='White' & exp$costume=='Poor'])
wa_a <- mean(exp$Mill_tax_signed[exp$race_actor=='White' & exp$costume=='Affluent'])
wa_dim <- wa_a - wa_p

## Black actor difference in means by costume
ba_p <- mean(exp$Mill_tax_signed[exp$race_actor=='Black' & exp$costume=='Poor'])
ba_a <- mean(exp$Mill_tax_signed[exp$race_actor=='Black' & exp$costume=='Affluent'])
ba_dim <- ba_a - ba_p

## Difference in difference
ddim <- wa_dim - ba_dim
ddim
```

```
## [1] 0.06730154
```

```
## Regression
lm(Mill_tax_signed ~ costume + blackactor + costume:blackactor, data=exp)
```

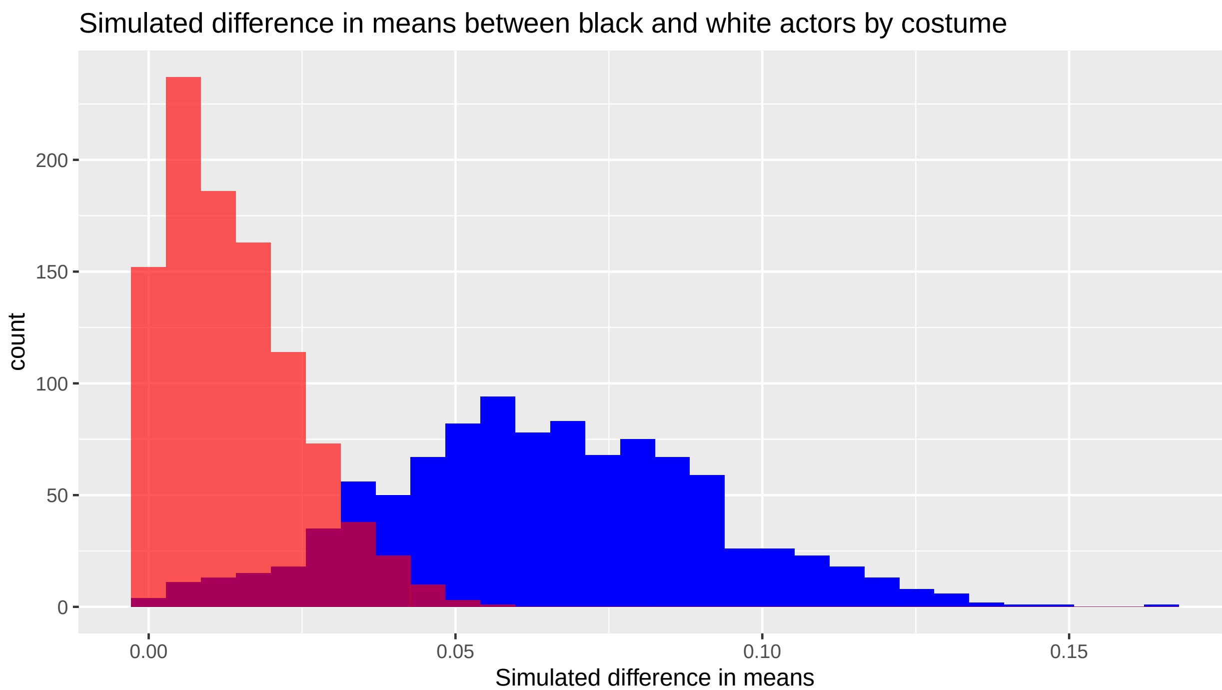
```
##
## Call:
## lm(formula = Mill_tax_signed ~ costume + blackactor + costume:blackactor,
##     data = exp)
##
```

```
## Coefficients:
##           (Intercept)            costumePoor             blackactor
##               0.11940               -0.06535               -0.04719
## costumePoor:blackactor
##               0.06730
```

The difference in means for white actors was found by subtracting the mean proportion of signers who saw a poor actor by the proportion of signers who saw an affluent actor, conditioned by the actor being white The same was done for black actors. The difference in means based on costume for each race was then used to calculate a new difference in means, which was found to be .067. In terms of the regression model, the mean proportion of signers for the millionaire tax was found to be .119, as represented by the intercept in the regression coefficients. However, when the actor was wearing a 'poor' costume, the estimated mean proportion of signatures given decreases by .065. When the actor was black, it decreases by .047. The interaction term, costumePoor:blackactor, is important for understanding the relationship between the two factors. Thus, the factors may be dependent upon each other when making estimations. We also can see that the value of the interaction term, .067, is the same is the difference in differences.

4. **Generating 1000 bootstrap replications of the difference in difference in means**

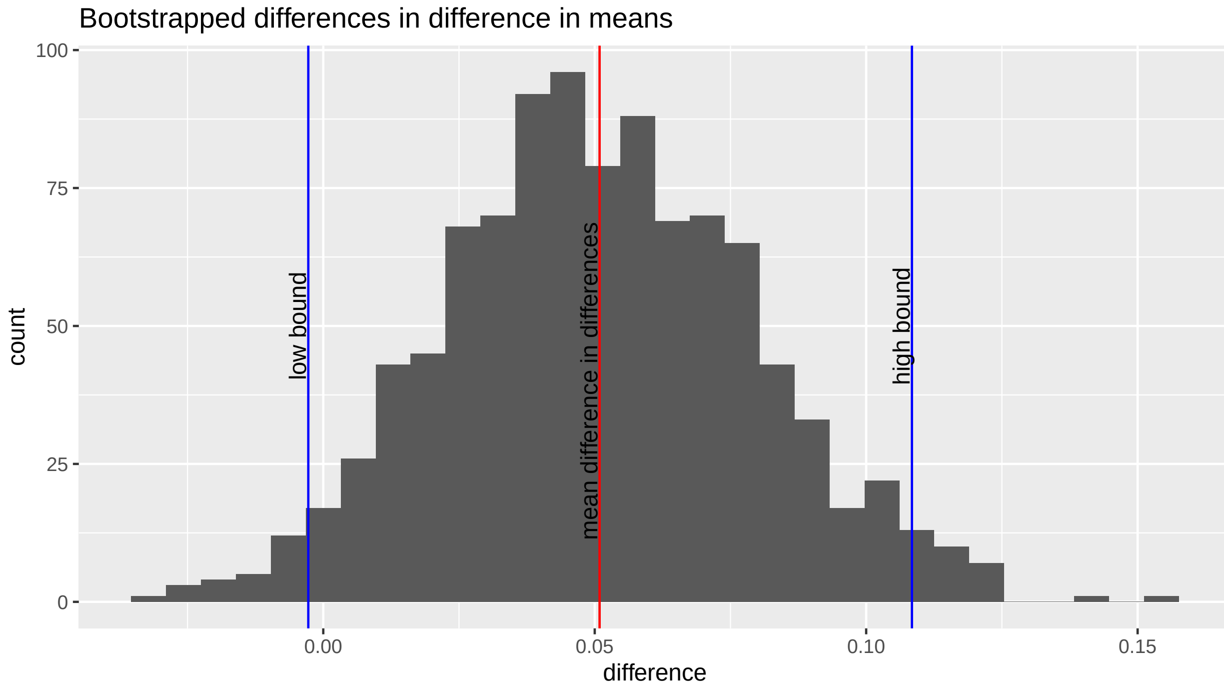
```r
## Bootstrap replications
w_actors <- c()
b_actors <- c()
for (i in 1:1000) {
  row_nums <- sample(1:nrow(exp), nrow(exp), replace=T)
  samp_exp <- exp[row_nums,]
  samp_wa_p <- mean(samp_exp$Mill_tax_signed[samp_exp$race_actor=='White'
                                             & samp_exp$costume=='Poor'])
  samp_wa_a <- mean(samp_exp$Mill_tax_signed[samp_exp$race_actor=='White'
                                             & samp_exp$costume=='Affluent'])
  samp_ba_p <- mean(samp_exp$Mill_tax_signed[samp_exp$race_actor=='Black'
                                             & samp_exp$costume=='Poor'])
  samp_ba_a <- mean(samp_exp$Mill_tax_signed[samp_exp$race_actor=='Black'
                                             & samp_exp$costume=='Affluent'])
  w_actors <- append(w_actors, abs(samp_wa_a - samp_wa_p))
  b_actors <- append(b_actors, abs(samp_ba_a - samp_ba_p))
}
samp_ddim <- w_actors - b_actors

## Confidence interval bounds
wa_bounds <- quantile(w_actors, c(.025, .975))
ba_bounds <- quantile(b_actors, c(.025, .975))
ddim_bounds <- quantile(samp_ddim, c(.025, .975))

## Bootstrap distribution 1
ggplot() +
  geom_histogram(aes(w_actors), fill='blue') +
  geom_histogram(aes(b_actors), fill='red', alpha=.65) +
  ggtitle('Simulated difference in means between black and white actors by costume') +
  xlab('Simulated difference in means')
```



There is little overlap, thus the distributions are centered around different values. It is important to consider the differing characteristics of each distribution. If a simulated difference in means for each race falls somewhere in the overlap, it could be interpreted that there is not much difference, or interaction, between the factors. Thus, it is important to run many simulations and analyze reference distributions as it is clear that there are quite notable differences.

For one, there is more spread in the distribution representing white actors compared to black actors. The difference in means for white actors has a peak around .07, but seems like it could easily fluctuate as there is not a strong peak. There is a clear, taller peak in difference in means for black actors at about .01. Secondly, the reference distribution for black actors seems as though it includes 0 in the confidence interval. Thus, for black actors, the decision for people to sign the millionaire tax petition was not heavily influence by the costume factor. The costume factor, did however, clearly have an affect on those who saw the white actor. Again, there is more spread, but the distribution hardly includes 0 at all in its range and very much likely does not have it within a 95% confidence interval.

```
## Bootstrap distribution 2
ggplot() +
  geom_histogram(aes(samp_ddim)) +
  geom_vline(xintercept=mean(samp_ddim), col='red') +
  geom_vline(xintercept=ddim_bounds, col='blue') +
  ggtitle('Bootstrapped differences in difference in means') +
  xlab('difference') +
  ylab('count') +
  annotate('text', x=ddim_bounds[1] - .002, y=50, label='low bound', angle=90) +
  annotate('text', x=ddim_bounds[2] - .002, y=50, label='high bound', angle=90) +
  annotate('text', x=mean(samp_ddim) - .002, y=40, label = 'mean difference in differences', angle=90)
```



The range of the reference distribution here does contain 0, but not within the confidence interval. Thus, there should be some sort of interaction present between the costume and black actor variables in 95% of simulations. Knowing this, it is important to compare the variables as did for the previous graph where black actors and white actors were represented by two different reference distributions.