Yichen Dong HW 9

Yichen Dong

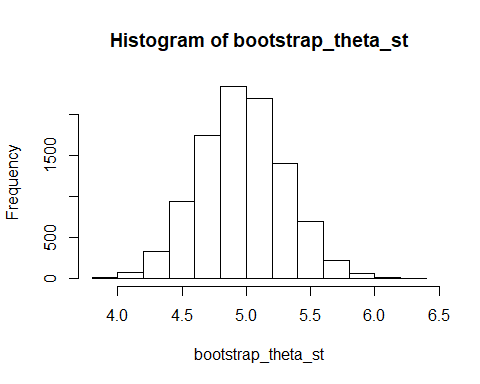
November 1, 2018

## Problem 1

stomach = c(25,42,45,46,51,103,124,146,340,396,412,876,1112)  
breast = c(24,40,719,727,791,1166,1235,1581,1804,3460,3808)  
log\_stomach = log(stomach)  
log\_breast = log(breast)  
mean\_log\_stomach = mean(log\_stomach)  
mean\_log\_breast = mean(log\_breast)  
  
b = 10000  
bootstrap\_theta\_st = NULL  
for(i in 1:b){  
 n = length(stomach)  
 data\_bootstrap = sample(log\_stomach,size = n, replace = TRUE)  
 bootstrap\_mean = mean(data\_bootstrap)  
 bootstrap\_theta\_st[i] = bootstrap\_mean  
}  
bootstrap\_theta\_bar\_st = mean(bootstrap\_theta\_st)  
bootstrap\_variance\_st = 1/(b-1)\*sum((bootstrap\_theta\_st-bootstrap\_theta\_bar\_st)^2)  
#Finding the CI using an alpha of .05 and using the Percentile Method  
sorted\_bootstrap\_theta\_st = sort(bootstrap\_theta\_st)  
bootstrap\_theta\_CI\_st = c(bootstrap\_theta\_bar\_st- sqrt(bootstrap\_variance\_st)\*qnorm(.975),bootstrap\_theta\_bar\_st +sqrt(bootstrap\_variance\_st)\*qnorm(.975))  
bootstrap\_theta\_CI\_st\_pct = c(sorted\_bootstrap\_theta\_st[round(b\*.025)],sorted\_bootstrap\_theta\_st[round(b\*.975)])  
bootstrap\_variance\_st

## [1] 0.1088757

hist(bootstrap\_theta\_st)



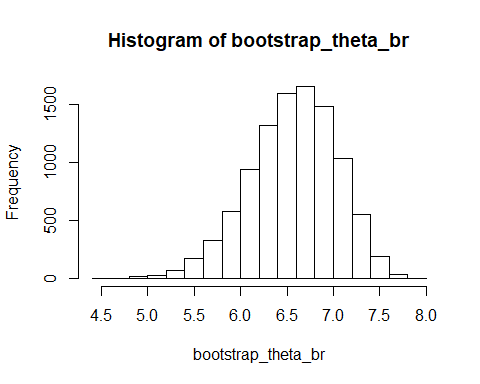
bootstrap\_theta\_CI\_st

## [1] 4.319724 5.613156

bootstrap\_theta\_CI\_st\_pct

## [1] 4.330936 5.623129

#For breasts  
b = 10000  
bootstrap\_theta\_br = NULL  
for(i in 1:b){  
 n = length(breast)  
 data\_bootstrap = sample(log\_breast,size = n, replace = TRUE)  
 bootstrap\_mean = mean(data\_bootstrap)  
 bootstrap\_theta\_br[i] = bootstrap\_mean  
}  
bootstrap\_theta\_bar\_br = mean(bootstrap\_theta\_br)  
bootstrap\_variance\_br = 1/(b-1)\*sum((bootstrap\_theta\_br-bootstrap\_theta\_bar\_br)^2)  
#Finding the CI using an alpha of .05 and using the Percentile Method  
sorted\_bootstrap\_theta\_br = sort(bootstrap\_theta\_br)  
bootstrap\_theta\_CI\_br = c(bootstrap\_theta\_bar\_br- sqrt(bootstrap\_variance\_br)\*qnorm(.975),bootstrap\_theta\_bar\_br +sqrt(bootstrap\_variance\_br)\*qnorm(.975))  
bootstrap\_theta\_CI\_br\_pct = c(sorted\_bootstrap\_theta\_br[round(b\*.025)],sorted\_bootstrap\_theta\_br[round(b\*.975)])  
hist(bootstrap\_theta\_br)



bootstrap\_variance\_br

## [1] 0.2191473

bootstrap\_theta\_CI\_br

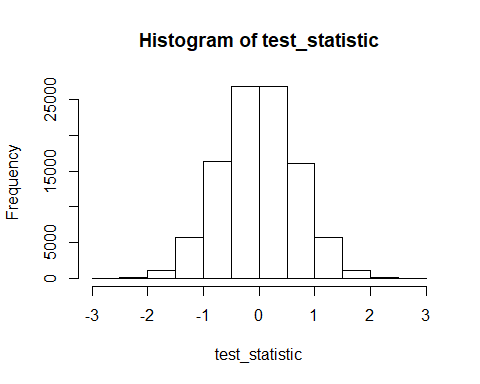
## [1] 5.647315 7.482358

bootstrap\_theta\_CI\_br\_pct

## [1] 5.562023 7.387759

We can see that the two confidence intervals are pretty close. ### Part b

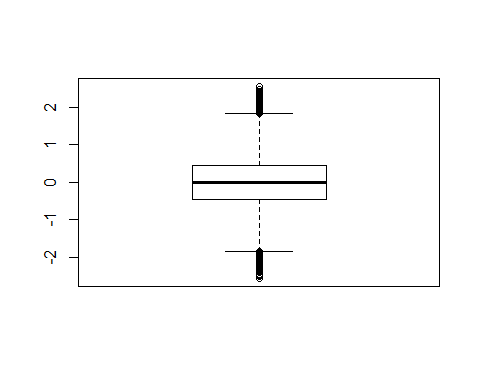
# I'm not sure how to permute, so I'm just going to mix up the variables many many times and hope that's a good approximation  
combined = c(log\_stomach,log\_breast)  
itr = 100000  
test\_statistic = NULL  
for(i in 1:itr){  
 sample = sample(combined)  
 permute\_stomach = sample[1:13]  
 permute\_breast = sample[14:24]  
 test\_statistic[i] = mean(permute\_stomach) - mean(permute\_breast)  
}  
hist(test\_statistic)



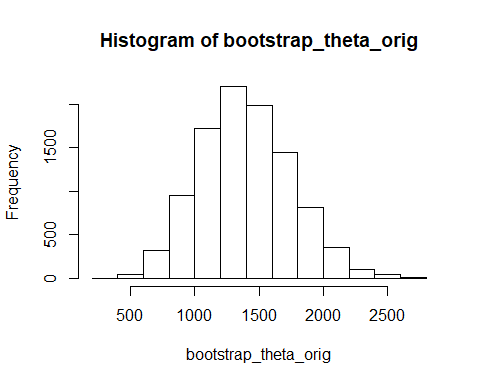
mean(test\_statistic)

## [1] -0.00214056

boxplot(test\_statistic)

 We can see that the different between the two groups is centered around 0. ### Part C

exp\_breast\_ci = exp(bootstrap\_theta\_CI\_br\_pct)  
b = 10000  
bootstrap\_theta\_orig = NULL  
for(i in 1:b){  
 n = length(breast)  
 data\_bootstrap = sample(breast,size = n, replace = TRUE)  
 bootstrap\_mean = mean(data\_bootstrap)  
 bootstrap\_theta\_orig[i] = bootstrap\_mean  
}  
bootstrap\_theta\_bar\_orig = mean(bootstrap\_theta\_orig)  
bootstrap\_variance\_orig = 1/(b-1)\*sum((bootstrap\_theta\_orig-bootstrap\_theta\_bar\_orig)^2)  
hist(bootstrap\_theta\_orig)



#Finding the CI using an alpha of .05 and using the Percentile Method  
sorted\_bootstrap\_theta\_orig = sort(bootstrap\_theta\_orig)  
bootstrap\_theta\_CI\_orig = c(bootstrap\_theta\_bar\_orig- sqrt(bootstrap\_variance\_orig)\*qnorm(.975),bootstrap\_theta\_bar\_orig +sqrt(bootstrap\_variance\_orig)\*qnorm(.975))  
bootstrap\_theta\_CI\_orig\_pct = c(sorted\_bootstrap\_theta\_orig[round(b\*.025)],sorted\_bootstrap\_theta\_orig[round(b\*.975)])  
exp\_breast\_ci

## [1] 260.349 1616.080

bootstrap\_variance\_orig

## [1] 126378.9

bootstrap\_theta\_CI\_orig

## [1] 699.9172 2093.4441

bootstrap\_theta\_CI\_orig\_pct

## [1] 746.5455 2122.9091

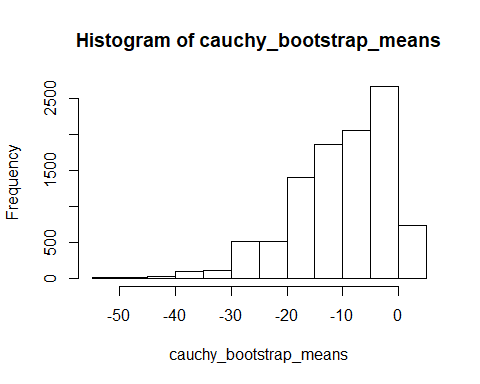
We can see that the confidence intervals for the exponential confidence interval is a lot different from the one for just the original data.

## Problem 2

cauchy = rcauchy(1000)  
mean(cauchy)

## [1] -10.24537

b = 10000  
cauchy\_bootstrap\_means = NULL  
for(i in 1:b){  
 n = length(cauchy)  
 bootstrap\_cauchy = sample(cauchy,size = n,replace = TRUE)  
 cauchy\_bootstrap\_means[i] = mean(bootstrap\_cauchy)  
}  
hist(cauchy\_bootstrap\_means)



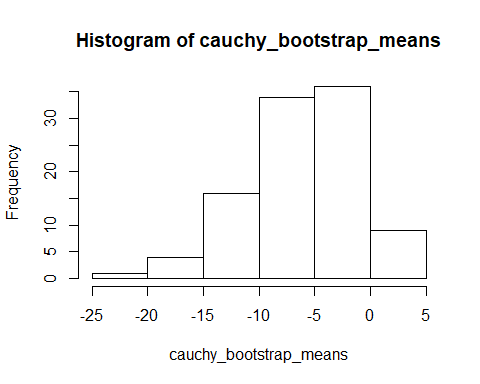
mean(cauchy\_bootstrap\_means)

## [1] -10.33495

#testing a smaller bootstrap  
cauchy = rcauchy(100)  
mean(cauchy)

## [1] -5.394489

b = 100  
cauchy\_bootstrap\_means = NULL  
for(i in 1:b){  
 n = length(cauchy)  
 bootstrap\_cauchy = sample(cauchy,size = n,replace = TRUE)  
 cauchy\_bootstrap\_means[i] = mean(bootstrap\_cauchy)  
}  
hist(cauchy\_bootstrap\_means)



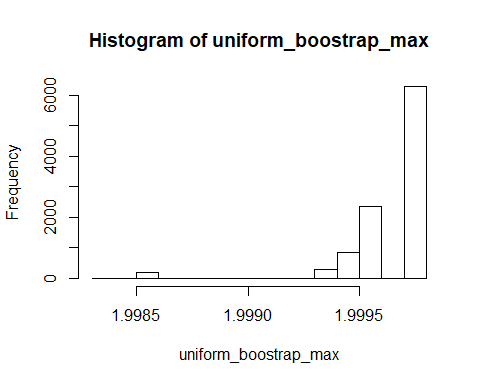
mean(cauchy\_bootstrap\_means)

## [1] -5.215819

# for estimating max  
uniform = runif(10000,0,2)  
max(uniform)

## [1] 1.99973

b = 10000  
uniform\_boostrap\_max = NULL  
for(i in 1:b){  
 n = length(uniform)  
 bootstrap\_uniform = sample(uniform,size=n,replace= TRUE)  
 uniform\_boostrap\_max[i] = max(bootstrap\_uniform)  
}  
hist(uniform\_boostrap\_max)



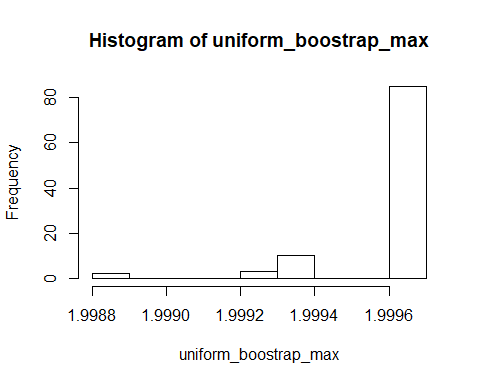
max(uniform\_boostrap\_max)

## [1] 1.99973

#testing a smaller bootstrap  
uniform = runif(10000,0,2)  
max(uniform)

## [1] 1.999659

b = 100  
uniform\_boostrap\_max = NULL  
for(i in 1:b){  
 n = length(uniform)  
 bootstrap\_uniform = sample(uniform,size=n,replace= TRUE)  
 uniform\_boostrap\_max[i] = max(bootstrap\_uniform)  
}  
hist(uniform\_boostrap\_max)

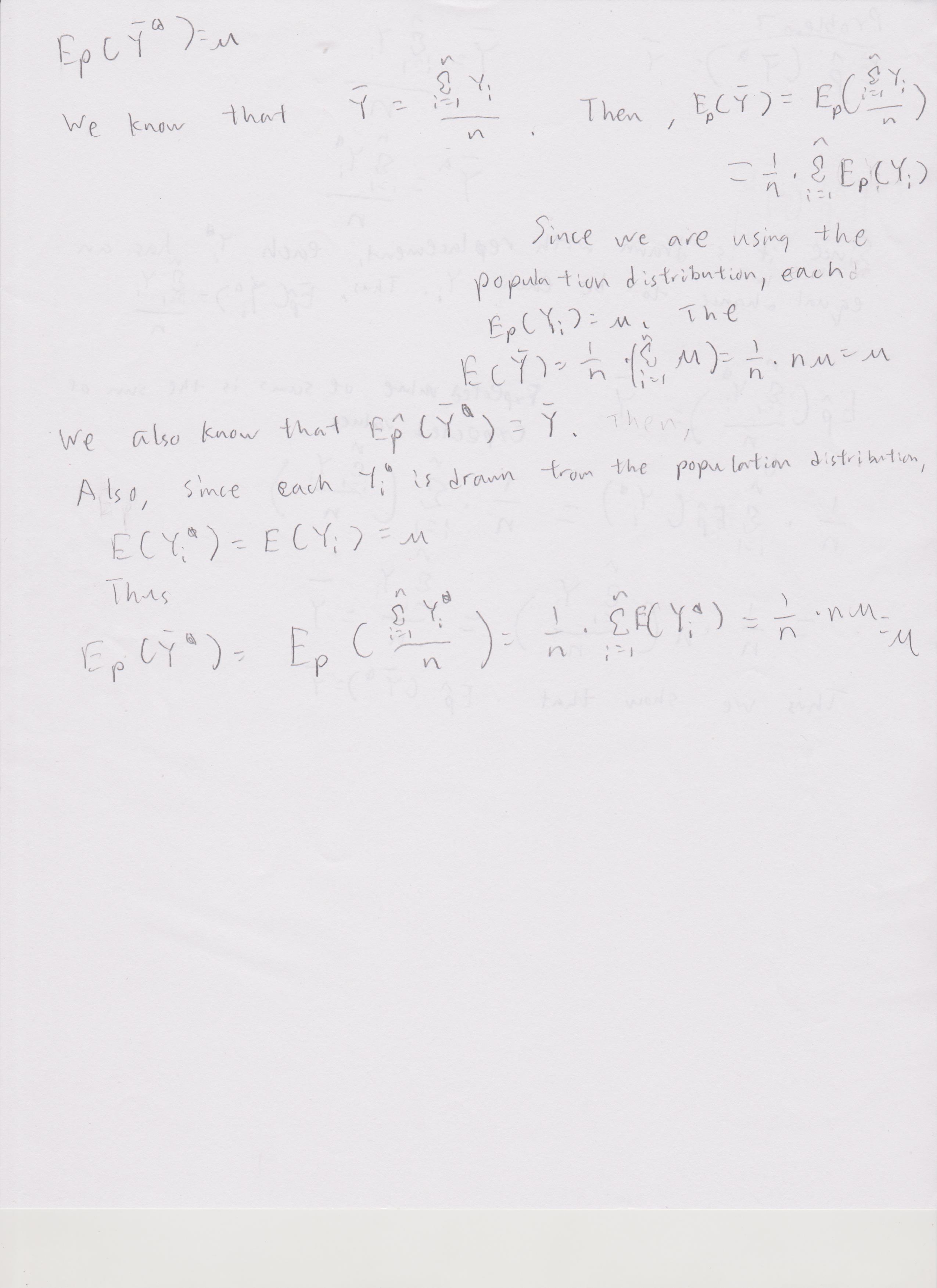
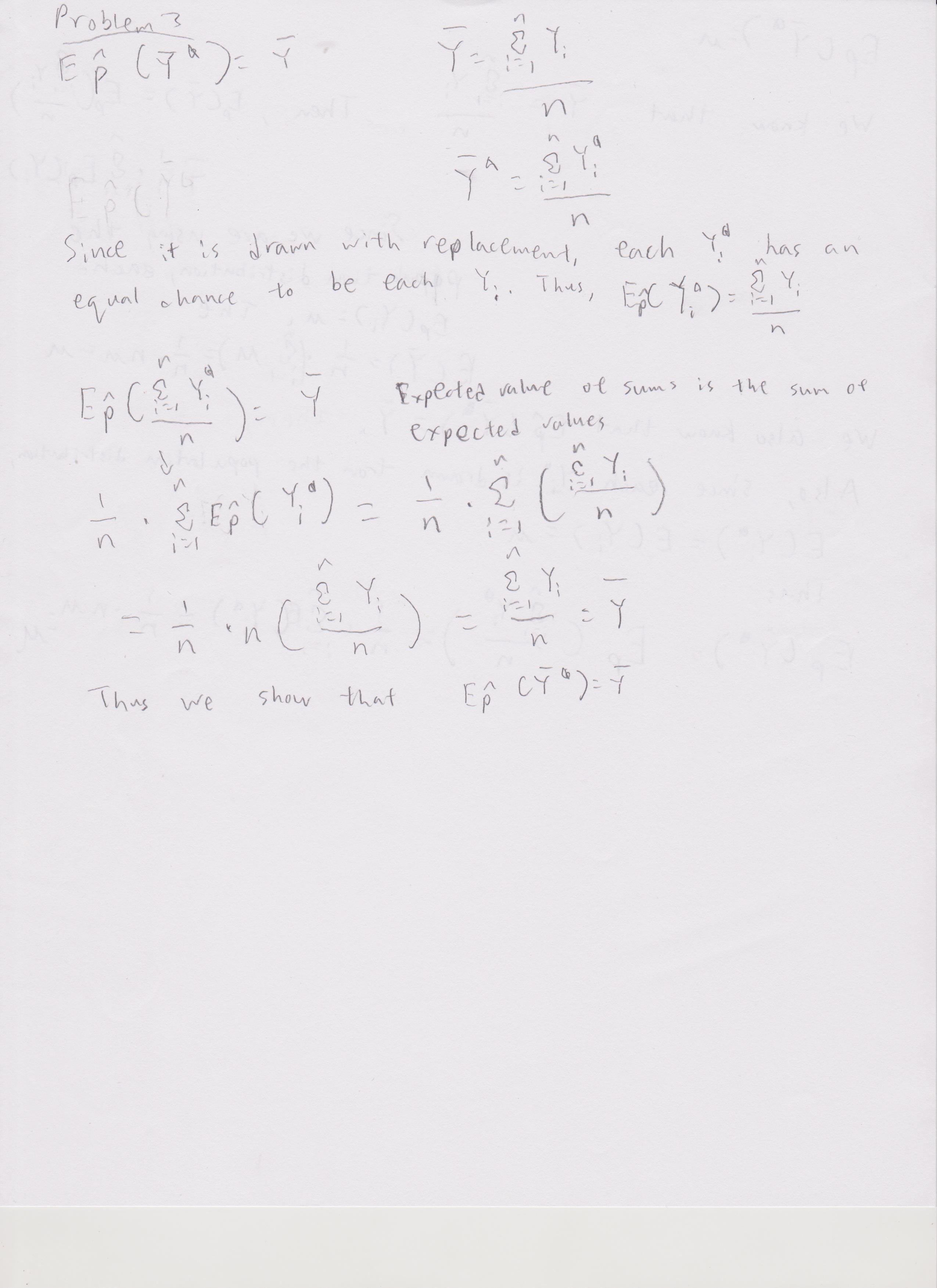


max(uniform\_boostrap\_max)

## [1] 1.999659

I’m honestly not too sure about this. I feel like they are predicting pretty accurately for the max and the mean of the cauchy. Maybe it’s because the histogram does not look centered around a single value, so that the variance of a small number of bootstraps could be very high?

## Problem 3



## Problem 4

p4\_rnorm = rnorm(100,0,1)  
mean(p4\_rnorm)

## [1] -0.1307336

##standard bootstrap  
b = 10  
p4\_boot\_mean = NULL  
for(i in 1:b){  
 n = length(p4\_rnorm)  
 bootstrap\_norm = sample(p4\_rnorm,size=n,replace=TRUE)  
 p4\_boot\_mean[i] = mean(bootstrap\_norm)  
}  
p4\_theta\_bar\_star = mean(p4\_boot\_mean)  
bias\_corrected = 2\*mean(p4\_rnorm) - mean(p4\_boot\_mean)  
p4\_b\_est\_variance = 1/(b-1)\*sum((p4\_boot\_mean - p4\_theta\_bar\_star)^2)  
p4\_theta\_bar\_star

## [1] -0.07590083

bias\_corrected

## [1] -0.1855664

p4\_b\_est\_variance

## [1] 0.007953236

##balanced bootstrap  
b=10  
balance\_p4\_rnorm = NULL  
for(i in 1:b){  
 balance\_p4\_rnorm = c(balance\_p4\_rnorm,p4\_rnorm)  
}  
balance\_p4\_rnorm\_permute = sample(balance\_p4\_rnorm, length(balance\_p4\_rnorm))  
balance\_p4\_means = NULL  
for(i in 1:b){  
 start = i\*100 -99  
 end = i\*100  
 balance\_p4\_means[i] = mean(balance\_p4\_rnorm\_permute[start:end])  
}  
p4\_balance\_theta\_bar\_star = mean(balance\_p4\_means)  
balance\_bias\_corrected = 2\*mean(p4\_rnorm) - mean(balance\_p4\_means)  
p4\_balance\_variance = 1/(b-1)\*sum((balance\_p4\_means-p4\_balance\_theta\_bar\_star)^2)  
mean(balance\_p4\_means)

## [1] -0.1307336

balance\_bias\_corrected

## [1] -0.1307336

p4\_balance\_variance

## [1] 0.005972435

We can see that the balanced bootstrap gave us an answer that was a lot closer to the real mean, with a smaller variance than the original bootstrap.