STAT 641: BOOTSTRAPPING METHODS

Jiyoun Myung

Department of Statistics and Biostatistics
California State University, East Bay

Spring 2021, Day 6

Your turn

A high school student was curious about the total number of minutes devoted to commercials during any given half-hour time period on basic and extended cable TV channels (B. Rodgers and T. Robinson, personal communication).

ID	Times	Cable
1	7.0	Basic
2	10.0	Basic
3	10.6	Basic
4	10.2	Basic
5	8.6	Basic
6	7.6	Basic
7	8.2	Basic
8	10.4	Basic
9	11.0	Basic
10	8.5	Basic

ID	Times	Cable
11	3.4	Extended
12	7.8	Extended
13	9.4	Extended
14	4.7	Extended
15	5.4	Extended
16	7.6	Extended
17	5.0	Extended
18	8.0	Extended
19	7.8	Extended
20	9.6	Extended

Your turn

```r
x.b <- c(7.0, 10.0, 10.6, 10.2, 8.6, 7.6,  8.2, 10.4, 11.0, 8.5)
x.e <- c(3.4, 7.8, 9.4, 4.7, 5.4, 7.6, 5.0, 8.0, 7.8, 9.6)
```

Question: Is there a significant difference in the length of commercials (in minutes) during random half hour periods? Use the **sample()** function.

- Bootstrap the difference in mean times, plot the distribution, and give summary statistics of the bootstrap distribution. Obtain a 95% bootstrap percentile confidence interval, and interpret this interval.
- What is the bootstrap estimate of the bias? What faction of the bootstrap standard error does this represent?
- Use the **boot()** function.

Answer

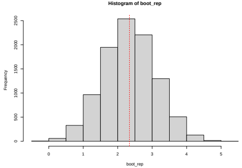
```r
set.seed(1)
x.b <- c(7.0, 10.0, 10.6, 10.2, 8.6, 7.6,  8.2,
         10.4, 11.0, 8.5)
x.e <- c(3.4, 7.8, 9.4, 4.7, 5.4, 7.6, 5.0, 8.0,
         7.8, 9.6)

boot_rep <- rep(0,10000)
for(i in 1:10000)
{
  boot.x <- sample(x.b, replace = T)
  boot.y <- sample(x.e, replace = T)
  boot_rep[i] <- mean(boot.x) - mean(boot.y)
}

# bootstrap distribution
(obs <- mean(x.b) - mean(x.e)) # observed value
```

```
## [1] 2.34
```

```r
hist(boot_rep)
abline(v = obs, col = "red", lty = 2)
```




```r
# bootstrap statistic
mean(boot_rep) # bootstrap mean
```

```
## [1] 2.344121
```

```r
sd(boot_rep) # bootstrap se
```

```
## [1] 0.759423
```

```r
quantile(boot_rep, c(.025, .975)) # 95% percentile CI
```

```
##  2.5% 97.5%
##  0.86  3.83
```

We are 95% confident that on average, the length of commercials on basic channels is from 0.86 to 3.83 min longer than on extended cable.

```r
# bootstrap estimate of bias
(bias <- mean(boot_rep) - obs)
```

```
## [1] 0.004121
```

```r
bias/sd(boot_rep)
```

```
## [1] 0.005426488
```

Bias is about 0.5% of the standard error.

boot package

Commercial time data

```
library(boot)
set.seed(123)
x.b <- c(7.0, 10.0, 10.6, 10.2, 8.6, 7.6,  8.2, 10.4, 11.0, 8.5)
x.e <- c(3.4, 7.8, 9.4, 4.7, 5.4, 7.6, 5.0, 8.0, 7.8, 9.6)

total <- c(x.b, x.e)
id <- as.factor(c(rep("1", length(x.b)), rep("2", length(x.e))))
total <- cbind(id, total)

meanDiff <- function(data, ind){
  y <- tapply(data[ind, 2], data[ind, 1], mean)
  y[1] - y[2]
}

b <- boot(total, meanDiff, strata = id, R = 10000)
mean(b$t) - b$t0 # bootstrap estimate of bias
```

```
##        1
## 0.012786
```

```
sd(b$t)
```

```
## [1] 0.7608422
```

```
n <- length(x.b); m <- length(x.e)
sqrt(var(x.b)/n + var(x.e)/m)
```

```
## [1] 0.7981228
```

```
sqrt((n-1)*var(x.b)/n^2 + (m-1)*var(x.e)/m^2)
```

```
## [1] 0.7571658
```

boot package

Mouse Data

```r
trt <- c(94, 197, 16, 38, 99, 141, 23)
ctrl <- c(52, 104, 146, 10, 51, 30, 40, 27, 46)

total <- c(trt, ctrl)
id <- as.factor(c(rep("x", length(trt)), rep("y", length(ctrl))))
total <- cbind(id, total)

b <- boot(total, meanDiff, strata = id, R = 10000)
mean(b$t) - b$t0 # bootstrap estimate of bias
```

```
##         1
## 0.2567254
```

```r
sd(b$t)
```

```
## [1] 26.85041
```

```r
n <- length(trt); m <- length(ctrl)
sqrt(var(trt)/n + var(ctrl)/m)
```

```
## [1] 28.93607
```

```r
sqrt((n-1)*var(trt)/n^2 + (m-1)*var(ctrl)/m^2)
```

```
## [1] 26.90811
```

```r
# hist(b$t)
```

You can also use “infer” or “rasmple” package for this case.