Business Analytics - ETC3250 2018 - Lab 8

The bootstrap

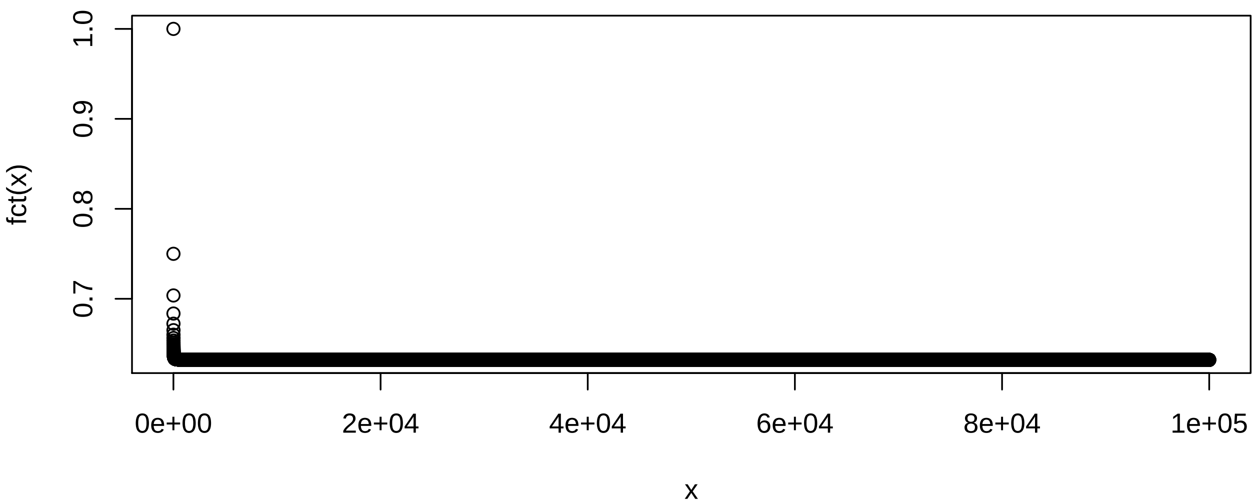
Souhaib Ben Taieb

19 April 2018

Exercise 1

Do the exercise 2 in Section 5.4 of ISLR.

- $\frac{n-1}{n}$
- $\frac{n-1}{n}$
- probability that the jth observation is not in the bootstrap sample = probability that the jth observation is not in the ith position where $i = 1, \ldots, n$ = (probability that the jth observation is not in the ith position)^n and probability that the jth observation is not in the ith position is given above.
- probability that the jth observation is in the bootstrap sample = 1 - probability that the jth observation is not in the bootstrap sample $= 1 - (1 - 1/5)^5 = 1 - (4/5)^5 = 67.2\%$
- $1 - (99/100)^{100} = 63.4\%$
- $1 - (1 - 1/10000)^{10000} = 63.2\%$



- We can clearly see the convergence to $1 - 1/e = 63.2\%$

```r
store <- rep (NA , 10000)
for (i in 1:10000) {
  store[i]  <- sum(sample (1:100 , rep =TRUE) == 4) > 0
}
mean(store)
# [1] 0.6319
```

The same conclusion as above.

Bootstrap confidence interval of the correlation coefficient

We will find a 95% confidence interval for the correlation coefficient of Median House value and average number of rooms in the Boston data set from the `MASS` package.

The functions `cor` and `cor.test` will compute the correlation and an asymptotic 95% confidence interval for it. This interval is based on Fisher's z transform

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

which is approximately normally distributed with variance $1/(n-3)$ where n is the number of observations. So if z_L and z_U are upper and lower limits for z, then

$$r_L = \frac{\exp(2z_L)-1}{\exp(2z_L)+1} \qquad \text{and} \qquad r_U = \frac{\exp(2z_U)-1}{\exp(2z_U)+1}$$

are upper and lower limits for r.

We will use the bootstrap to test if this is a good approximation in this case.

Exercise 2

Check that the confidence interval returned by `cor.test` is computed using the above transformation.

```
library(MASS)

n <- nrow(Boston)
r <- cor(Boston$medv, Boston$rm)

# Fisher interval
cor.test(Boston$medv, Boston$rm)
#
#   Pearson's product-moment correlation
#
# data:  Boston$medv and Boston$rm
# t = 21.722, df = 504, p-value < 2.2e-16
# alternative hypothesis: true correlation is not equal to 0
# 95 percent confidence interval:
#  0.6474346 0.7378075
# sample estimates:
#       cor
# 0.6953599

z <- 0.5*log((1+r)/(1-r))
zint <- z + 1.96/sqrt(n-3)*c(-1,1)
rint <- (exp(2*zint)-1)/(exp(2*zint)+1)
print(rint)
# [1] 0.6474337 0.7378082
```

Exercise 3

Compute a 95% bootstrap confidence interval for the correlation. You will need to sample rows of the `Boston` matrix.

```
B <- 1000
rb <- numeric(B)
for(i in 1:B)
{
  bootstrapdata <- Boston[sample(n, replace=TRUE),]
```

```
  rb[i] <- cor(bootstrapdata$medv, bootstrapdata$rm)
}
quantile(rb, prob=c(0.025,0.975))
#      2.5%     97.5%
# 0.6139358 0.7668311
```

Exercise 4

Write a function that will return a bootstrap confidence interval for the correlation of any two numeric variables of the same length. Your function should take four arguments:

- `x`: a numeric vector of data
- `y`: a numeric vector of data
- `level`: the probability coverage of the confidence interval with default value of 0.95
- `B`: the number of bootstrap samples with default value of 1000.

```
bootstrap.cor.int <- function(x, y, level=0.95, B=1000)
{
  n <- length(x)
  rb <- numeric(B)
  for(i in 1:B)
  {
    j <- sample(n, replace=TRUE)
    rb[i] <- cor(x[j],y[j])
  }
  alpha = 1-level
  return(quantile(rb, prob=c(alpha/2, 1-alpha/2)))
}

bootstrap.cor.int(Boston$medv,Boston$rm,B=10000)
#      2.5%     97.5%
# 0.6098922 0.7669034
```