STAT403 HW6

Chongyi Xu

May 10, 2018

Question 1

Assume that your data consists of $x_1, \cdots, x_n$, n values. When we generate the bootstrap sample, we sample with replacement of these n points to obtain a set of IID new points $x_1^*, \cdots, x_n^*$ such that

$$P(x_l^* = x_1) = P(x_l^* = x_2) = \cdots = P(x_l^* = x_n) = \frac{1}{n}$$

for each l. This new dataset is called a bootstrap sample.

(a) Show that the bootstrap sample is an IID random sample from $\hat{F}_n$, where

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$$

Denote Y to be a random variable that has $\frac{1}{n}$ probability with $x_1, \cdots, x_n$. Therefore, $P(Y \leq y) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq y)$, which is the same equation above. In the other words, Y is as same as bootstrap sample of x

(b) Assume we want to use the bootstrap to estimate the variance of the sample mean. It is well-known that the variance of sample mean can be approximated by the sample variance divided by n, the sample size. Let $\bar{X}_n^* = \frac{1}{n}\sum_i X_i^*$ be the sample mean of a bootstrap sample. Show that

$$Var(\bar{X}_n^*) = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n^2}S^2 n$$

Remark $S^2 n = \frac{1}{n-1}\sum_{i=1}(x_i - \bar{x})^2$

$$\begin{aligned}
Var(\bar{X}_n^*) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i^*) \\
&= \frac{1}{n^2} \cdot n(E(X_i^{*2}) - E(X_i^*)^2) \\
&= \frac{1}{n}(E(X_i^{*2}) - E(X_i^*)^2) \\
&= \frac{1}{n}(\frac{1}{n}\sum x_i^2 - \bar{x}^2) \\
&= \frac{1}{n^2}\sum(x_i^2 - \bar{x}^2) \\
&= \frac{1}{n^2}\sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{n-1}{n^2}S^2 n
\end{aligned}$$

Question 2

In this problem, we will use the bootstrap technique to analyze the faithful dataset. We focus on the standard deviation of the variable waiting.

(a) Apply the bootstrap 10000 times to show the bootstrap distribution of the SD.

```
library(ggplot2)
```

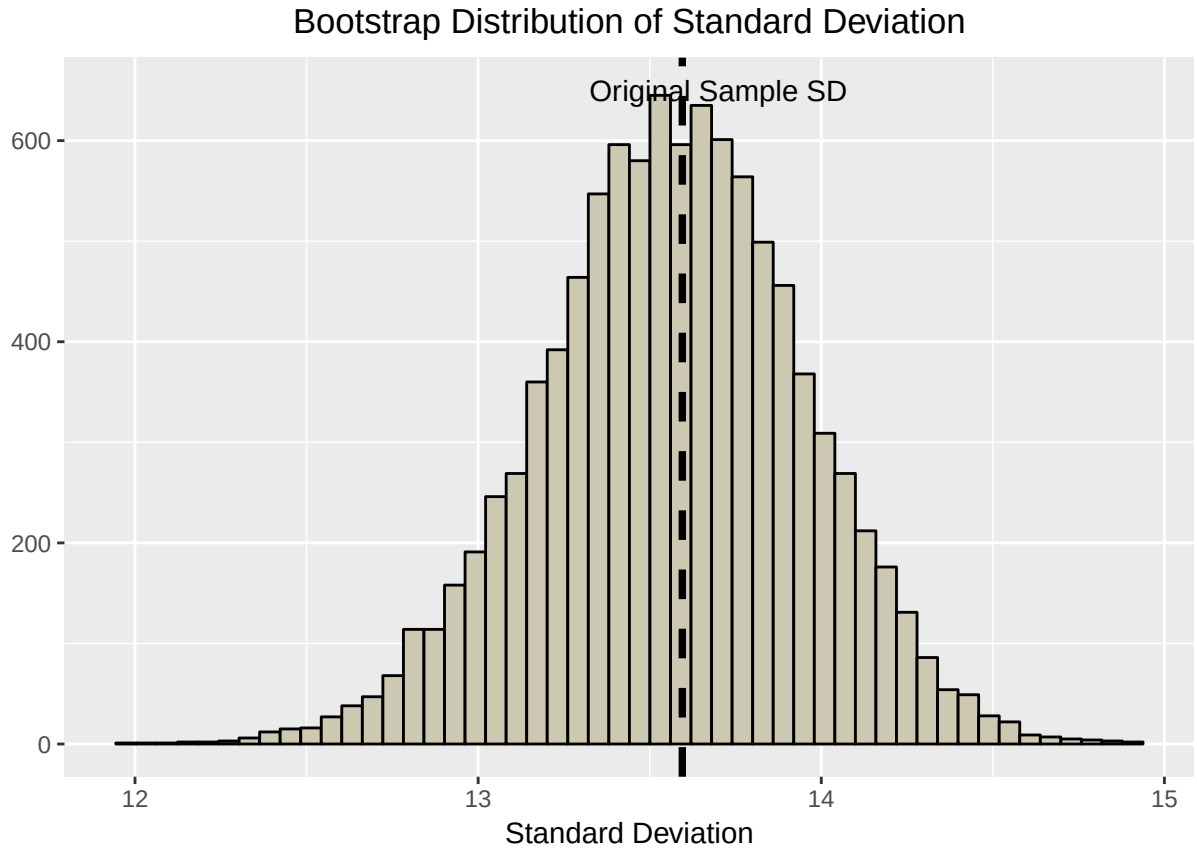
```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
dat <- faithful$waiting
sd_dat <- sd(dat)

N <- 10000
boot_var <- rep(0, N)

for (i in 1:N) {
  boot_dat <- sample(dat, size=length(dat), replace=TRUE)
  boot_var[i] <- sd(boot_dat)
}

ggplot() + geom_histogram(aes(boot_var),
                          bins=50,
                          color="black",
                          fill="cornsilk3") +
  geom_vline(xintercept=sd_dat, size=1.3,
             linetype='dashed')   +
  annotate('text', x=13.7,
           y=650, label='Original Sample SD') +
  xlab('Standard Deviation') + ylab('') +
  ggtitle('Bootstrap Distribution of Standard Deviation') +
  theme(plot.title = element_text(hjust = 0.5))
```



(b) What are the bootstrap estimate of the variance and MSE of the sample SD?

```
print(paste('The boostrap estimate of the variance is ', var(boot_var)))
```

```
## [1] "The boostrap estimate of the variance is  0.147264009555532"
```

```
print(paste('The boostrap estimate of the MSE is ', mean((boot_var-sd_dat)^2)))
```

```
## [1] "The boostrap estimate of the MSE is  0.148251470503007"
```

(c) Use both the asymptotic normality method and the quantile method to construct 95% CIs of the SD.

```
CI <- c(sd_dat-qnorm(0.975)*sd(boot_var), sd_dat+qnorm(0.975)*sd(boot_var))
print(paste('The 95% CI using asymptotic normality is ', CI[1], ',', CI[2]))
```

```
## [1] "The 95% CI using asymptotic normality is  12.8428377461364 , 14.3471098338624"
```

```
CI <- quantile(boot_var, c(0.025,0.975))
print(paste('The 95% CI using quantile ', CI[1], ',', CI[2]))
```

```
## [1] "The 95% CI using quantile  12.7873672485149 , 14.2919091735301"
```

(d) Let σ be the true SD of variable `waiting`. Assume we want to test the hypothesis:

$$H_0 : \sigma = 15$$

using the bootstrap sample. There are many ways to test this hypothesis using the bootstrap approach. Here we simply use the bootstrap variance estimate to test the hypothesis. What is the p-value?

```r
print(paste('The p-value is ', t.test(boot_var, rep(15,N))$p.value, ', which means we reject the null hy
```

```
## [1] "The p-value is  0 , which means we reject the null hypothesis"
```

(e) Briefly explain the benefit of increasing the number of bootstrap samples.

Since bootstrap is a Monte Carlo method, its Monte Carlo error will decrease as the sample size increases, which will improve the estimation.

Question 3

In this problem, we will use the bootstrap to analyze the odds ratio of UC-Berkeley's admission dataset. In particular, we will focus on the department A.

(a) Use the bootstrap to compute the MSE of the odds ratio OR.

```r
table <- UCBAdmissions[,,1]
# Mark 12=Male got admitted, 22=Female got admitted, 11=Male got rejected, 21=Female got rejected.
dat <- c(rep(12,table[1,1]), rep(22, table[1,2]), rep(11, table[2,1]), rep(21, table[2,2]))
or <- (length(dat[dat==12])*length(dat[dat==21]))/(length(dat[dat==22])*length(dat[dat==11]))

N <- 10000
boot_or <- rep(0, N)

for (i in 1:N) {
  boot_dat <- sample(dat, size=length(dat), replace=TRUE)
  boot_or[i] <- (length(boot_dat[boot_dat==12])*length(boot_dat[boot_dat==21]))/(length(boot_dat[boot_da
}

print(paste('The MSE is ', mean((boot_or-or)^2)))
```

```
## [1] "The MSE is  0.00883845855343066"
```

(b) If there is no gender bias, the odds ratio will be 1. Use the bootstrap to compute the p-value of testing

$$H_0 : \text{no gender bias in this contingency table.}$$

```r
print(paste('The p-value is ', t.test(boot_var, rep(1,N))$p.value, ', which means we reject the null hyp
```

```
## [1] "The p-value is  0 , which means we reject the null hypothesis"
```

(c) In this case, the parametric bootstrap and the empirical bootstrap are the same procedure. Explain why.

The reason could be that in this problem, the data can be treated as categorical which will lead parametric boostrap and empirical bootstrap to the same procedure.