Bootstrap method

Woojoo Lee

We want to assess the uncertainty of a statistic under the nonparametric framework.

$\rightarrow$ The bootstrap makes us do it ! i.e. we can estimate variance of a statistic and compute the confidence interval with bootstrap.

Let $T_n = \bar{X}_n$.

Q) How can you get the variance estimate for $\bar{X}_n$?

Remark) The key problem is that we do not know how to generate random samples from F.

Suppose that we have $x_1, \cdots, x_n$ from an unknown distribution F. The parameter of interest is

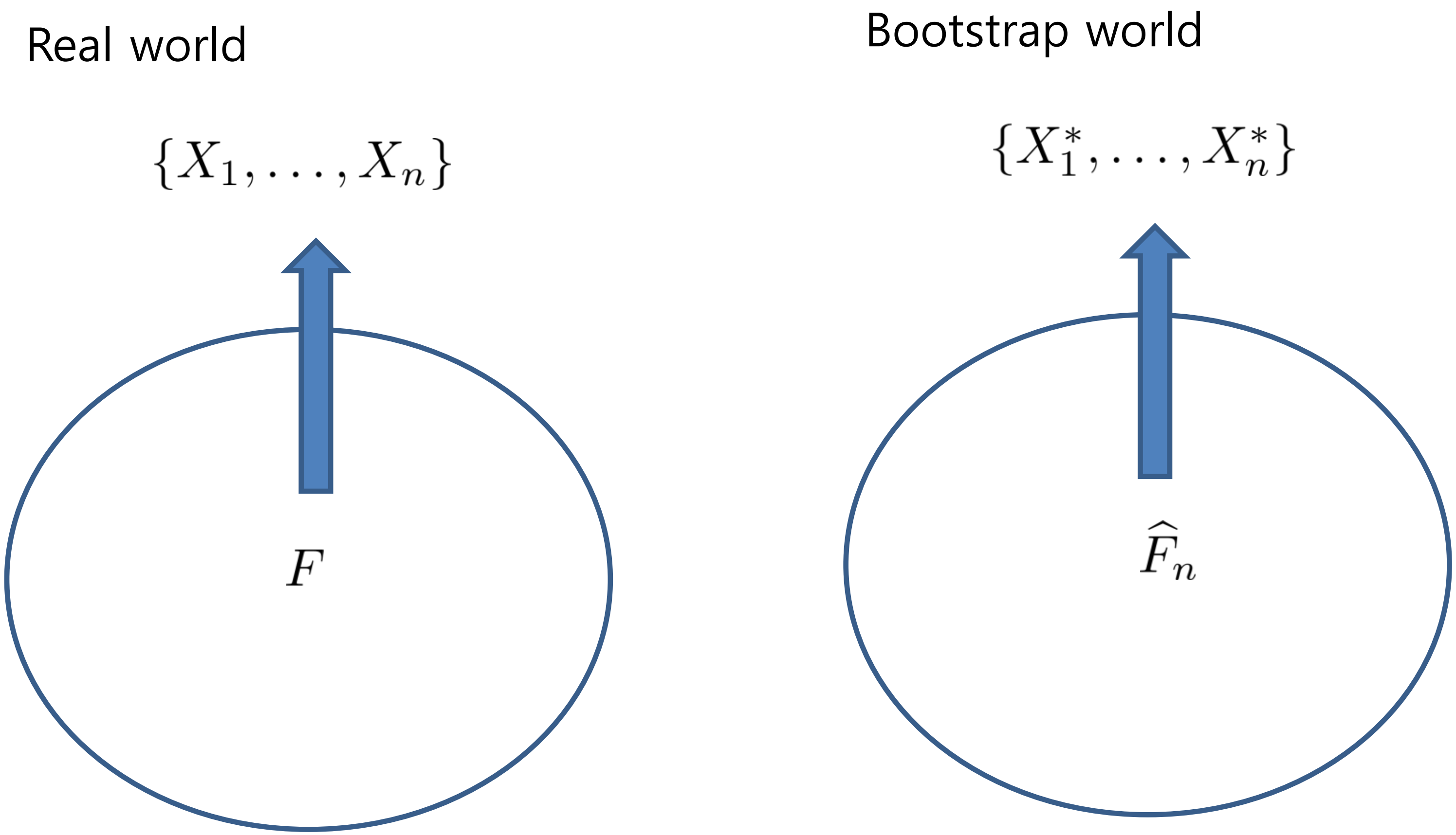
$$\theta = E(h(X)) = \int h(x) dF(x)$$

We note that

- $F \approx \hat{F}_n$
- $\hat{\theta} \approx \int h(x) d\hat{F}_n(x)$
- Often, the dificulty lies in computing var$(\hat{\theta})$.

Var$(\hat{\theta})$ requires "many samples" ! but we have only one sample in practice.

$\rightarrow$ Pretend to know F by using $\hat{F}_n$ and generate new samples from $\hat{F}_n$!



Key procedure: Sampling n observations from $\widehat{F}_n$ with replacement !

Bootstrap algorithm for computing the variance of a statistic

1. Compute the empirical CDF $\hat{F}_n$
2. Sample n observations from $\hat{F}_n$ with replacement
3. Compute the statistic of interest with the bootstrap samples.
4. Repeat the above steps 2-3 B times.

Denote the statistic from bth bootstrap sample by $\hat{\theta}_b$.
Compute

$$var_B(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\theta}_B)^2$$

where

$$\bar{\theta}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b.$$

Somtimes, we are interested in the bias of an estimator. The bias is defined as

$$E(\hat{\theta}) - \theta.$$

We can estimate this bias from bootstrap:

$$\frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b - \hat{\theta}$$

.

R-example

```
### Example: calculating se for median
set.seed(1201)
data<-rnorm(100,5,3)

B<-1000
b.samples<-lapply(1:B,function(i) sample(data,replace=T))
b.median<-sapply(b.samples,median)


hist(b.median)

sqrt(var(b.median))

## in theory,
## p*(1-p)/(n*f(5)^2)=1/(100*4*f(5)^2)=0.1414   (f(5)=dnorm(0,0,3))
## sqrt(0.1414)=0.376
```

R-example

```
### Example: calculating se for skewness

library(moments)
set.seed(1201)
data<-rnorm(100,5,3)

B<-1000
b.samples<-lapply(1:B,function(i) sample(data,replace=T))
b.skewness<-sapply(b.samples,skewness)


hist(b.skewness)

sqrt(var(b.skewness))
```

R-example

```
library(boot)
data(bigcity)

## we want to know the mean ratio of the populations,
## i.e. pop 1930/pop 1920

row.bigcity<-dim(bigcity)[1]

boots.bigcity<-function(index){
            b.bigcity<-bigcity[index,]
            b.ratio<-sum(b.bigcity$x)/sum(b.bigcity$u)
            return(b.ratio)
}

B<-1000
b.samples<-lapply(1:B,function(i) sample(c(1:row.bigcity),replace=T))
b.ratio<-sapply(b.samples,boots.bigcity)

hist(b.ratio)

sqrt(var(b.ratio))
```

R-example

```
#### As an alternative, you may use "boot".
#### Before calling boot,
#### you need to define a function that will return the statistic
#### that you want to bootstrap.

library(boot)
ratio <- function(d, indices) sum(d$x[indices])/sum(d$u[indices])
RES.city<-boot(bigcity, ratio, R = 999)

boot.ci(RES.city,type=c("norm","basic","perc","bca"))
```

Bootstrap confidence intervals

Bootstrap confidence intervals

We will study three different bootstrap confidence intervals.

The normal confidence interval based on bootstrap:

Suppose that $\hat{\theta}$ is the observed statistic. Then, the $1-\alpha$ normal interval is given by

$$(\hat{\theta} - z_{\alpha/2}\hat{se}_{boot}, \hat{\theta} + z_{1-\alpha/2}\hat{se}_{boot})$$

where $\widehat{se}_{boot}$ is the bootstrap estimate of standard error.

Remark) This interval is not accurate if the sampling distribution of $\hat{\theta}$ is not close to normal.

Suppose that $\widehat{\theta^*}_\alpha$ is α-percentile of $\widehat{\theta^*}_i$. The $1-\alpha$ bootstrap percentile interval is given by

$$(\widehat{\theta^*}_{\alpha/2}, \widehat{\theta^*}_{1-\alpha/2})$$

Some properties

- Easy to use !
- This works well when the bootstrap distribution is symmetric and centered on the observed statistic.
- Many literatures report that this can be narrow in small samples.

The $1-\alpha$ basic bootstrap confidence interval is given by

$$(2\hat{\theta} - \widehat{\theta^*}_{1-\alpha/2}, 2\hat{\theta} - \widehat{\theta^*}_{\alpha/2})$$

Q) Derive the basic bootstrap CI.

Remark) The key idea is that the distribution of $\widehat{\theta}^* - \widehat{\theta}$ is approximately the same as that of $\widehat{\theta} - \theta$.

Algorithm for bootstrap-t (studentized) confidence interval

- Generate B bootstrap samples ($*$ denotes the statistic from the bootstrap samples.)

- Compute $t_i^* = \frac{\hat{\theta_i^*} - \hat{\theta}}{\hat{\sigma_i^*}}$ where $\hat{\theta}_i^*$ and $\hat{\sigma_i^*}$ denote the statistic and s.e. from ith bootstrap sample, respectively.

- Compute α-percentile of t_i^*: find t_α^* satisfying $\#(t_i^* \leq t_\alpha^*)/B = \alpha$.

- The bootstrap-t $1 - \alpha$ confidence interval is given by

$$(\hat{\theta} - t^*_{\alpha/2}\hat{\sigma}, \hat{\theta} + t^*_{1-\alpha/2}\hat{\sigma})$$

Some properties

- The bootstrap t-confidence interval reflects the skewness of the data.

- More accurate than the percentile and the basic confidence intervals.

R-example for various bootstrap CIs

```
### Example: calculating se for median
set.seed(1201)
data<-rnorm(100,5,3)

B<-1000
b.samples<-lapply(1:B,function(i) sample(data,replace=T))
b.median<-sapply(b.samples,median)


hist(b.median)

sqrt(var(b.median))

### normal interval
c(mean(b.median)-2*sqrt(var(b.median)),mean(b.median)+2*sqrt(var(b.median)))

### percentile interval
c(quantile(b.median,0.025),quantile(b.median,0.975))

### basic interval
c(2*median(data)-quantile(b.median,0.975),2*median(data)-quantile(b.median,0.025))
```

Bootstrap in regression

Consider a simple linear regression model:

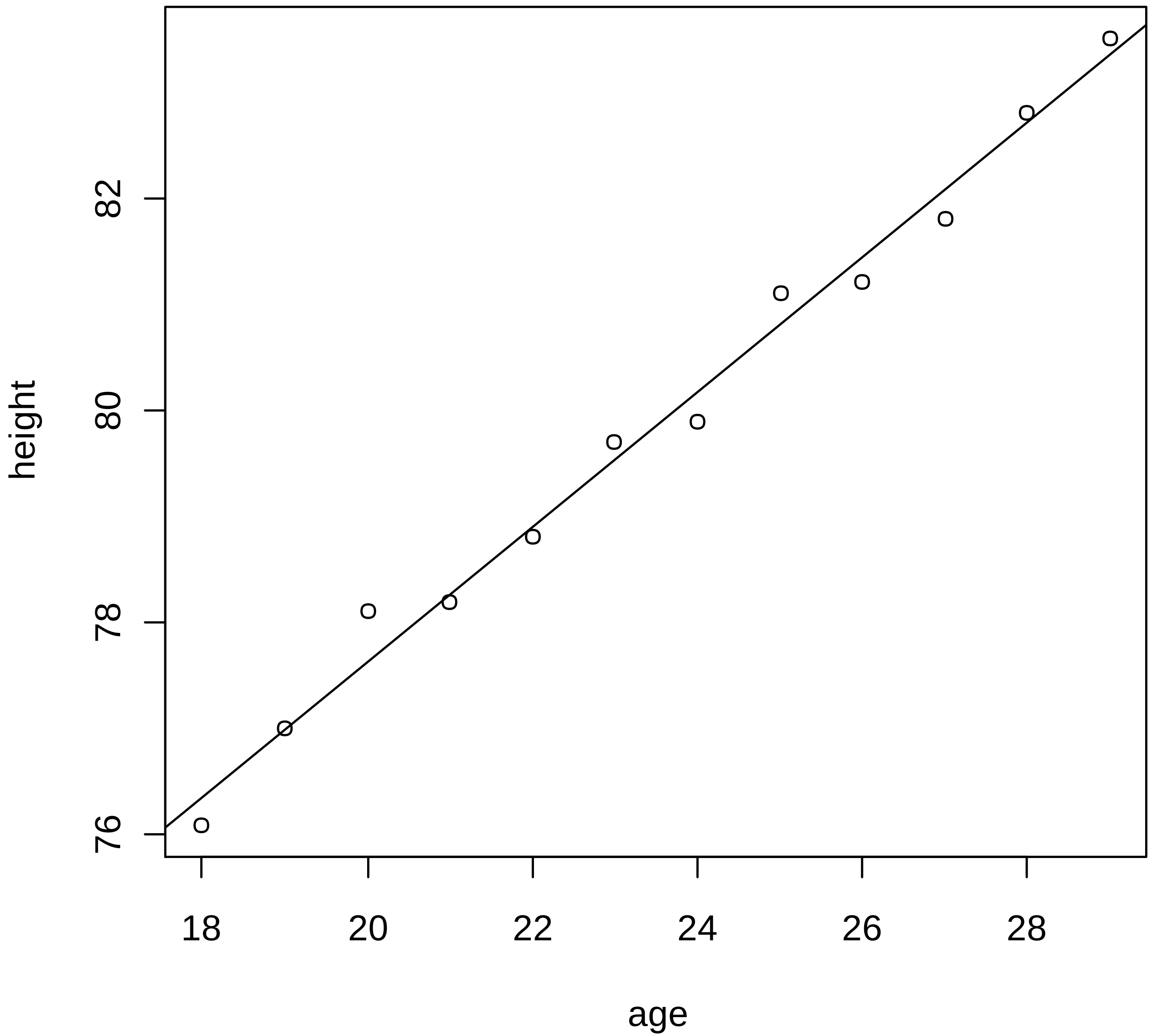
$$y_i = \beta_0 + x_i\beta_1 + e_i$$

where e_i is i.i.d. with mean 0 and variance σ^2. Note that there is no parametric assumption on the distribution for e_i.

Two alternatives bootstrapping methods are available.

- case 1) bootstrapping pairs (x_i, y_i)
- case 2) bootstrapping residuals

R-example for bootstrapping in regression



Q) Explain the bootstrap method for case 1 (case sampling).

```
boots.pair<-function(index){
        b.age<-age[index]
        b.height<-height[index]
        b.coeff<-coef(lm(b.height~b.age))[2]
        return(b.coeff)
}

B<-1000
set.seed(1210)
b.samples<-lapply(1:B,function(i)
sample(c(1:length(height)),replace=T))
b.PWD<-sapply(b.samples,boots.pair)

mean(b.PWD)
sd(b.PWD)
```

Q) Explain the bootstrap method for case 2 (model-based sampling).

```
lm.res<- lm(height ~ age)
residual<-resid(lm.res)
fit<-fitted(lm.res)

boots.resid<-function(index){
        newy<-fit+residual[index]
        b.coeff<-coef(lm(newy~age))[2]
        return(b.coeff)
}

B<-1000
set.seed(1210)
b.samples<-lapply(1:B,function(i)
sample(c(1:length(height)),replace=T))
b.PWD<-sapply(b.samples,boots.resid)

mean(b.PWD)
sd(b.PWD)
```

Q) Which one is better ?