# Introducció Bootstrap. Sessio 5 (Estimació variancia estimador. Estimació Biaix)

Jordi Ocaña, Sergi Civit,

2 de maig de 2018

## 1 Estimació Bootstrap de la variancia d'un estimador

Considerem el problema d'estimar el paràmetre 'theta' igual a la ratio o quocient de les seves, desconegudes, variàncies poblacionals

$$\theta = sigma_1^2/sigma_2^2$$

Ara farem servir el bootstrap per **calcular la variància d'un estimador**.

Per estimar **theta** un **estimador molt raonable** és el quocient de les respectives variàncies mostrals no esbiaixades,

$$var(x_1)/var(x_2)$$

Pel mètode bootstrap el procés és:

```
> # Considerem aquestes dades "reals":
> x1 = c(4.44, 2.88, 0.91, 0.44, 0.15, 5.70, 2.95, 1.05, 1.63, 1.39, 0.38, 0.33)
> x2 = c(7.66, 27.85, 20.67, 4.50, 23.38, 7.83, 36.29, 29.38, 38.30, 15.93, 13.16,
+        4.30, 53.23, 30.02, 15.63, 2.80, 11.24)
> n1 = length(x1)  # 12
> n2 = length(x2)  # 17
> # Considerem el problema d'estimar 'theta'
>
> # De vegades utilitzem el bootstrap per estimar alguna característica de
> # la distribució d'un estadístic o un estimador com ara la seva variància.
>
> # Per estimar theta un podria pensar que un estimador molt raonable és el
> # quocient de les respectives variàncies mostrals no esbiaixades, var(x1)/var(x2).
>
> # De vegades interessa estimar la variància de var(x1)/var(x2)
> # com estimador de theta = sigma1^2/sigma2^2.
>
```

```
> # Generem mitjançant bootstrap no paramètric B rèpliques d'aquest estimador:
> B = 10000
> set.seed(23771)
> theta.boots = replicate(B, {
+   x1.sim = sample(x1, replace = TRUE)
+   x2.sim = sample(x2, replace = TRUE)
+   var(x1.sim) / var(x2.sim)
+ })
> # Calculem la variància mostral d'aquests B valors:
> var(theta.boots)

[1] 0.0001108918
```

# 2 Biaix d'un estimador

També ens pot interessar **calcular el biaix d'un estimador**.

Per exemple: $var(x_1)$ és un estimador no esbiaixat de $\sigma_1^2$ $var(x_2)$ és un estimador no esbiaixat de $\sigma_2^2$

però:

$var(x_1)/var(x_2)$ HO ÉS DE $\theta = sigma_1^2/sigma_2^2$??

El **veritable biaix** seria

$$Evar(x_1)/var(x_2) - theta$$

el qual **el podem estimar** mitjançant la mitjana mostral dels B valors que acabem d'obtenir:

**theta.estim al "món bootstrap"juga el paper de theta al "món real", per tant** i per tant podem obtenir l'estimació bootstrap del biaix serà:

```
> # O també ens pot interessar el biaix d'un estimador.
> # Per exemple:
> #   var(x1) és un estimador no esbiaixat de sigma1^2
> #   var(x2) és un estimador no esbiaixat de sigma2^2
> # però:
> #   var(x1)/var(x2) ho és de theta = sigma1^2/sigma2^2 ???
>
> # El veritable biaix seria  E{var(x1)/var(x2)} - theta
> # E{var(x1)/var(x2)} l'estimem mitjançant la mitjana mostral dels B valors
> # que acabem d'obtenir:
> Eboot = mean(theta.boots)
> # theta.estim al "món bootstrap" juga el paper de theta al "món real",
> # per tant, l'estimació bootstrap del biaix serà:
> Eboot - var(x1)/var(x2)

[1] 0.001830584
```

```
> # O millor: en pura ortodoxia bootstrap seria s'hauria de restar el quocient
> # de les variàncies mostrals esbiaixades (raó?)
> theta.estim = (((n1 - 1) / n1) * var(x1)) / (((n2 - 1) / n2) * var(x2))
> theta.estim

[1] 0.01542598

> Eboot - theta.estim

[1] 0.002243043
```

# 3 Trimmed mean: Variància i Biaix. Enfoc Bootstrap i Jacknife

```
> # Aquesta mostra fa el paper d'unes "dades reals" però en realitat sabem que
> # procedeix d'una N(15, 3)
> # ATENCIÓ: AIXÒ ÉS AIXÍ PER QUE ESTEM EN UNA SITUACIÓ "DE LABORATORI", AMB
> # DADES REALS LÒGICAMENT DESCONEIXERÍEM COMPLETAMENT ELS PARÀMETRES REALS
>
> x <- c(15.54, 21.06, 16.52, 13.62, 16.14, 10.98, 13.53, 16.02, 16.79, 15.90)
> n <- length(x)
> # mitjana retallada de les dades "reals":
> mtrim <- mean(x, trim = 0.2)
> # valors jackknife, suprimint cada vegada un valor:
> mtrim_i <- numeric(length = n)
> for (i in 1:n) mtrim_i[i] <- mean(x[-i], trim = 0.2)
> # possiblement, una manera més eficient:
> mtrim_i = vapply(1:n, function(i) mean(x[-i], trim = 0.2), FUN.VALUE = 0.0)
> mtrim. <- mean(mtrim_i)
> # estimació jackknife de la variància de la mitjana retallada:
> varJ <- ((n - 1) / n) * sum((mtrim_i - mtrim.)^2)
> # estimació jackknife del biaix de la mitjana retallada:
> bJ <- (n - 1) * (mtrim. - mtrim)
> # mitjana retallada "corregida pel biaix" (mtrim - bJ):
> mtrimJ <- n * mtrim - (n - 1) * mtrim.
> # Enfoc bootstrap
>
> # una mostra aleatòria amb reemplaçament de x
> # (equivalent a dir: una mostra a partir de la distribució empírica Fn),
> # és a dir, una remostra bootstrap:
> sample(x, replace = TRUE)

 [1] 13.53 16.14 10.98 15.90 13.62 15.90 16.52 21.06 10.98 16.02

> # nombre de rèpliques bootstrap:
> B <- 10000
```

```
> # determinació la variància i el biaix de la mitjana retallada 20%
>
> # mitjana retallada d'una (re)mostra bootstrap:
> mean(sample(x, replace = TRUE), trim = 0.2)

[1] 15.94

> # B valors de la mitjana retallada a partir de B remostres bootstrap:
> mtrim.boots <- replicate(B, mean(sample(x, replace = TRUE), trim = 0.2))
> # les 10 primeres:
> mtrim.boots[1:10]

 [1] 16.09333 15.63667 14.80500 16.03167 15.32667 16.20667 16.10333 16.29500
 [9] 14.37000 15.10000

> # Estimació bootstrap (no paramètrica) de la variància de mitjana retallada:
> var(mtrim.boots)

[1] 0.5163037

> # Estimació bootstrap (no p.) del biaix
> mean(mtrim.boots) - mean(x)

[1] 0.003384333

> # o bé: ????
> mean(mtrim.boots) - mean(x, trim = 0.2)

[1] -0.009949

> # discutir...
>
>
> # Simulació per determinar la veritable variància i el veritable biaix
> # de la mitjana retallada del 0.2,
> # quan les dades procedeixen d'una normal de mitjana 15 i variància 9
>
> # simulem 1000000 de valors de la mitjana retallada:
> m <- 1000000
> mtrim.sims <-  replicate(m, mean(rnorm(10, mean = 15, sd = 3), trim = 0.20))
> # variància:
> varMtrim <- var(mtrim.sims)
> varMtrim

[1] 1.019578

> # i biaix:
> bMtrim <- mean(mtrim.sims) - 15
> bMtrim
```

```
[1] 0.0006802496

> # és nul el biaix?
> Lower<-bMtrim - (qnorm(0.975) * sqrt(varMtrim / m))
> Upper<-bMtrim + (qnorm(0.975) * sqrt(varMtrim / m))
> IC<-c(Lower, Upper)
> IC

[1] -0.001298808  0.002659307

> # La simulació no descarta el valor 0 de biaix
>
> # Bootstrap paramètric. Ara estem en una situació d'inferència paramètrica:
> # podem assumir que les dades procedeixen d'una distribució normal però
> # desconeixem els paràmetres d'aquesta distribució;
> # els hem d'estimar a # partir de la pròpia mostra
> var(replicate(B, mean(rnorm(n, mean = mean(x), sd = sd(x)), trim = 0.20)))

[1] 0.7869281

>
```