Introducció Bootstrap. Sessio 2 (NO Paramètric versus Paramètric)

Jordi Ocaña, Sergi Civit,

18 d'abril de 2018

1 Introducció: Situació de laboratori

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 COM A MOLT PODRÍEM TENIR ALGUNA IDEA DE LA SEVA FORMA per exemple que SEMBLA O S'ARPOXIMA A UNA NORMAL

2 Mostra

```
> # Lectura de les dades PROCEDENTS D'UNA NORMAL
> 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)
> mu <- 15
> sigma <- 3
```

3 Estudi de la distribució de l'estadístic t

A partir de la mostra anterior anaema dur a terme l'estudi bootstrap de la distribució QUE ENS PROPORCIONA LA METODOLOGIA PARAMÈTRICA
Estadístic t

```
> tStud <- function(x, mitjana)
+ {
+         (mean(x) - mitjana) / sqrt(var(x)/length(x))
+ }
```

3.1 Estudi de la distribució de l'estadístic t: Càlculs i gràfic

Obtenim:

- Mitjana mostral,
- Desviació típica mostral,
- Error estàndard de la mitjana mostral,
- Estimació de l'error estàndard de la mitjana mostral

```
> # Mitjana mostral, estimació de la veritable mitjana (que recordem que és 15)
> xBarra <- mean(x)
> xBarra

[1] 15.61

> # Desviació típica mostral, estimació de la veritable sigma (que és 3)
> s.x <- sqrt(var(x))
> s.x

[1] 2.62857

> # Veritable valor de l'error estàndard de la mitjana mostral:
> sigma.xBarra <- sigma/sqrt(n)
> sigma.xBarra

[1] 0.9486833

> # Estimació de l'error estàndard de la mitjana mostral:
> s.xBarra <- s.x/sqrt(n)
> s.xBarra

[1] 0.8312267

> # Calculem l'estadístic t i el grafiquem:
> t.x <- tStud(x, mu)
> t.x

[1] 0.7338552

> rang.t <- seq(from=-4, to=+4, by=0.1)
> dens.veritat <- dt(rang.t, df=n-1)
> windows(21,21)
> plot(rang.t, dens.veritat, type="l", col="green", ylim=c(0,0.4))
```

3.2 Estudi de la distribució de l'estadístic t: Valors crítics i Interval de confiança

Molts cops, es pot interessar coneixer certs "valors crítics"d'aquesta distribució mostral.

Aui tenim els **valors crítics segons la "veritable"distribució mostral de t**

```
> tCritStud <- qt(c(0.975, 0.025), df = n - 1)
> tCritStud
```

```
[1]  2.262157 -2.262157
```

Amb aquests valors podriem calcular un **interval de confiança paramètric** a partir **d'una t(n - 1)gl**

```
> xBarra - tCritStud * s.xBarra
```

```
[1] 13.72963 17.49037
```

Suposem ara la **Normal com a model aproximat** i el grafiquem el tabulem i calculem el interval de confiança

```
> dens.normAprox <- dnorm(rang.t)
> lines(rang.t, dens.normAprox, type="l", col="blue")
```

Obtenim **valors crítics segons aproximació normal**:

```
> tCritNorm = qnorm(c(0.975, 0.025))
> tCritNorm
```

```
[1]  1.959964 -1.959964
```

Realitzem **Interval de Confiança segons aproximació normal**:

```
> xBarra - tCritNorm * s.xBarra
```

```
[1] 13.98083 17.23917
```

4 Bootstrap NO PARAMÈTRIC de la distribució de l'estadístic t

```
> B <- 10000
> sample(x, replace = TRUE)
```

```
 [1] 16.14 13.62 21.06 13.62 21.06 21.06 16.02 13.62 15.54 13.53
```

```
> set.seed(127)
> mostres.bootstrap <- matrix(sample(x, replace=T, size=B*n), ncol=B)
> mostres.bootstrap[,1:10]
```

```
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] 16.52 16.79 16.14 16.79 16.52 21.06 15.90 16.14 16.52 10.98
 [2,] 15.54 13.62 16.14 15.90 15.54 13.62 13.53 21.06 16.52 16.02
 [3,] 16.52 16.02 16.02 16.14 15.90 15.54 16.79 13.62 10.98 10.98
 [4,] 21.06 21.06 15.90 16.52 21.06 16.02 16.14 15.54 13.62 15.90
 [5,] 13.62 16.14 10.98 16.52 16.79 21.06 16.52 13.62 16.79 16.52
 [6,] 16.02 13.62 13.62 15.90 16.02 13.62 13.62 16.52 16.79 15.54
 [7,] 10.98 16.02 13.62 21.06 13.53 13.53 15.54 15.90 13.62 15.54
 [8,] 16.52 21.06 16.14 15.90 10.98 16.02 21.06 16.14 16.79 21.06
 [9,] 16.02 10.98 16.14 13.62 15.54 15.90 16.14 21.06 13.62 16.52
[10,] 16.79 15.54 13.62 16.02 13.53 15.90 13.62 16.52 13.53 10.98
```

```
> t.bootstrap <- apply(mostres.bootstrap, 2, tStud, mitjana=xBarra)
> t.bootstrap[1:10]
```

```
 [1]  0.43437145  0.47895372 -1.37915517  1.41891456 -0.08298137  0.70798559
 [7]  0.39447074  1.23138657 -1.12476800 -0.59853481
```

```
> dens.bootstrap <- density(t.bootstrap,from=-4, to=+4)
> lines(dens.bootstrap, type="l", col="red")
```

4.1 Bootstrap NO PARAMÈTRIC de la distribució de l'estadístic t: Valors crítics i Interval de Confiança

Obtenim **valors crítics**

```
> tCritBoot = quantile(t.bootstrap, probs = c(0.975, 0.025))
```

Interval de confiança bootstrap no paramètric

```
> xBarra - tCritBoot * s.xBarra
```

```
   97.5%     2.5%
13.84873 17.46900
```

5 Bootstrap PARAMÈTRIC: Gràfic, Valors crítics, Interval de Confiança

Suposem que: EN REALITAT CREIEM de que la forma de la distribució es NORMAL

```
> # Si en realidad nos podemos fiar de que la forma de la distribución es normal
> # Bootstrap paramètric normal:
> # Una sola remostra:
> rnorm(n, mean=xBarra, sd=s.x)
```

```
 [1] 15.08031 16.34837 15.04077 14.52995 19.77730 17.58484 16.35833 13.93648
 [9] 15.36730 17.20246
```

```
> B <- 10000
> set.seed(127)
> mostres.bootstrap <- matrix( rnorm(B*n, mean=xBarra, sd=s.x), ncol=B)
> t.bootstrap.param <- apply(mostres.bootstrap ,2, tStud, mitjana=xBarra)
> dens.bootstrap.param <- density(t.bootstrap.param,from=-4, to=+4)
> lines(dens.bootstrap.param, type="l", col="brown")
```

Obtenim **valors crítics**

```
> # valors crítics segons bootstrap paramètric
> tCritBoot.param = quantile(t.bootstrap.param, probs = c(0.975, 0.025))
```

Interval de confiança bootstrap paramètric

```
> xBarra - tCritBoot.param * s.xBarra
```

```
   97.5%     2.5%
13.72557 17.44400
```

6 Càlcul d'una probabilitat P[t>valor]

Suposem que per decidir quin és el valor que a la seva dreta i deixa 0.025 de probabilitat, hem decidit utilitzar les taules de la normal (és a dir, hem considerat que **t té distribució N(0,1)**)

```
> z0.05 <- qnorm(0.025, lower.tail = FALSE)  # 1.959964
```

En teoria tindríem:

```
> pnorm(z0.05, lower.tail = FALSE)
```

```
[1] 0.025
```

6.1 Càlcul d'una probabilitat P[t>valor] t-student n-1

```
> # Però la veritable probabilitat és:
> pt(z0.05, df = n - 1, lower.tail = FALSE)
```

```
[1] 0.04082457
```

6.2 Càlcul d'una probabilitat P[t>valor] Bootstrap No paramètric

Si utilitzem l'aproximació a la veritable distribució de t que ens ha proporcionat el bootstrap no paramètric, aquesta probabilitat és:

```
> length(t.bootstrap[t.bootstrap > z0.05])/B
```

```
[1] 0.0321
```

6.3 Càlcul d'una probabilitat P[t>valor] Bootstrap Paramètric

Si utilitzem l'aproximació a la veritable distribució de t que ens ha proporcionat el bootstrap paramètric, aquesta probabilitat és:

```
> length(t.bootstrap.param[t.bootstrap.param > z0.05])/B
```

```
[1] 0.0406
```