Método Bootstrap

Los métodos Bootstrap son una clase de métodos no paramétricos de los métodos Monte Carlo que estiman la distribución de una población $\mathbf{x} = \{x_1, \ldots, x_N\}$ mediante remuestreo. Los métodos de remuestreo tratan una muestra observada como una población finita, y de ella se generan muestras aleatorias para estimar las características de la población y hacer inferencias sobre la población muestreada. Lo más importante de usar estos métodos es que la muestra es la única información disponible.
Sea $\mathbf{x} = \{x_1, \ldots, x_N\}$ una *m.o.* (muestra observada). La base de los métodos de remuestreo es reemplazar la distribución (original) de $\mathbf{x}$ por un modelo dado por la distribución empírica de los datos $\mathbf{x}$. Pero, ¿Cómo generamos la distribución empírica de los datos $\mathbf{x}$? La respuesta es el siguiente algoritmo.

Algoritmo 1. Distribución empírica de $\mathbf{x}$.

1. Dar $\mathbf{x} = \{x_1, \ldots, x_N\}$.
2. Generar $k \sim U\{1, \ldots, N\}$ y definir $X^* = x_k$.

Cómo k fue elegida al azar en el conjunto $\{1, \ldots, N\}$ entonces X^* hereda ésta aleatoriedad. A la distribución de X^* se le conoce como la *distribución empírica* de los datos $x_1, \ldots, x_N$. Así, del algoritmo 1 obtenemos $X_1^*, \ldots, X_N^*$ que se distribuyen (uniformente) en el conjunto $\mathbf{x}$.
En la literatura normalmente se denota como $F_N(x)$ a la distribución empírica, y se demuestra en inferencia estadística que $F_N(x)$ es un estimador de $F(x)$. Además, $F_N(x)$ es una estadística suficiente para $F(x)$. De este modo, toda la información de F que está contenida en $\mathbf{x}$ está contenida en F_N. De hecho, podemos decir entonces que F_N es una función de distribución que se distribuye uniformemente en el conjunto $\mathbf{x}$. Es decir, no

nos hagamos, F_N es la distribución de X^* (aunque nosotros lo denotaremos como P_X^*). Una de las aplicaciones más importantes de Bootstrap es estimar intervalos de confianza.

Ejemplo 1.

Supongamos que tenemos un estimador *plug-in* $\hat{\theta}$, es decir,

$$\theta = \hat{\theta}(X_1, \ldots, X_m),$$

donde $X_1, \ldots, X_m \sim P_\theta$ y queremos encontrar un intervalo de confianza, tal que

$$P_\theta(\hat{\theta} - a \leq \theta \leq \hat{\theta} + b) = 1 - \alpha. \tag{1}$$

En clases vimos que una manera de crear un intervalo de confianza usando muestras Bootstrap, es calcular el intervalo

$$P_X^*(\hat{\theta}^* - a \leq \theta(P_X^*) \leq \hat{\theta}^* + b),$$

en lugar del intervalo (1). Donde P_X^* es la distribución empírica de la muestra, $a = \theta^*_{1-\alpha/2} - \hat{\theta}$ y $b = \theta^*_{\alpha/2} - \hat{\theta}$.
Así, tenemos el siguiente algoritmo.

Algoritmo 2. Para estimar un intervalo de confianza para θ.

1. Dar una m.o. $x_1, \ldots, x_m$ y $\hat{\theta}$ un estimador *plug-in* de θ.
2. Generar, para $j = 1, \ldots, N$,
$$x^{*(j)} = \{x_{j_1}, \ldots, x_{j_m}\},$$
donde $j_1, \ldots, j_m \sim U\{1, \ldots, m\}$.
3. Hacer, para $j = 1, \ldots, N$
$$\hat{\theta}^{*(j)} = \hat{\theta}(x^{*(j)}).$$
4. Ordenar de menor a mayor la muestra $\{\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(N)}\}$. Digamos que la muestra ordenada es $\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(N)}$.
5. Hacer $l = \lceil \frac{\alpha}{2} N \rceil$ y $u = \lceil (1 - \frac{\alpha}{2}) N \rceil$.
6. Hacer $t = \hat{\theta}(x_1, \ldots, x_m)$
7. Y obtenemos el intervalo de confianza
$$\left(2t - \hat{\theta}^*_{(u)}, 2t - \hat{\theta}^*_{(l)}\right).$$

Hagamos un ejemplo.

Ejemplo 2. *Supongamos que $X_1, \ldots, X_m \sim N(0,1)$. Y queremos un intervalo de confianza para la media. Entonces, la implementación queda como sigue.*

```
#Primero vamos a calcular el intervalo de confianza con boostrap
SimpleBIC <- function(x,estimador.theta, N, alpha=0.05){
  #Generamos las muestras boostrap
  n<-length(x)
  thetas.gorros<-c()
```

```
  for (j in 1:N){
    X.bos<-sample(x, n, replace=TRUE)
    thetas.gorros[j]<-estimador.theta(X.bos)
  }
  #Construimos el intervalo de confianza
  t<-estimador.theta(x)
  l<-ceiling(0.5 * alpha * N)
  u<-ceiling((1-0.5 * alpha) * N)
  thetas.gorros.ordenados<-sort(thetas.gorros)
  return(c(2 * t -thetas.gorros.ordenados[u], 2 * t - thetas.
   gorros.ordenados[l]))
}
```

Luego, implementamos el intervalo exacto (no estoy seguro que "exacto" sea la palabra correcta). Tal intervalo es

$$\left(\bar{x} - \frac{t_{1-\alpha/2,n-1}\hat{\sigma}}{\sqrt{n}}, \bar{x} + \frac{t_{1-\alpha/2,n-1}\hat{\sigma}}{\sqrt{n}}\right). \tag{2}$$

```

#Luego calculamos el intervalo de confianza de manera exacta

ExactoIC <- function(x, alpha=0.05){
n <-length(x)
m <-mean(x)
s <-sd(x)
c <-qt(1 - 0.5*alpha, n-1)
return(c(m-c*s/sqrt(n), m+ c*s/sqrt(n)))
}
```

Y comparamos los intervalos, haciendo 100 repeticiones. Los resultados se pueden ver en la figura 2.

```
I1<-c()
I2<-c()
II1<-c()
II2<-c()
for (i in 1:100){
  x=rnorm(100)
  I1[i]=SimpleBIC(x,mean,1000,alpha = 0.05)[1]
  I2[i]=SimpleBIC(x,mean,1000,alpha = 0.05)[2]
  II1[i]=ExactoIC(x,0.05)[1]
  II2[i]=ExactoIC(x,0.05)[2]
}

C<-c()
for (j in 1:100){
C[j]=0
}

#Y graficamos

```

```
plot(C, type="l",col="limegreen",xlim=c(1,100),ylim = c(-1,1),ylab
    = "mean")
lines(I1, type="o", pch=22, lty=1, col="blue")
lines(I2, type="o", pch=22, lty=1, col="blue")
lines(II1, type="o", pch=22, lty=2, col="gray")
lines(II2, type="o", pch=22, lty=2, col="gray")
legend("topright",c("Intervalo Boot" , "Intervalo Ex"),lty=1,col=c
   ("blue","gray"),cex =0.5)
```

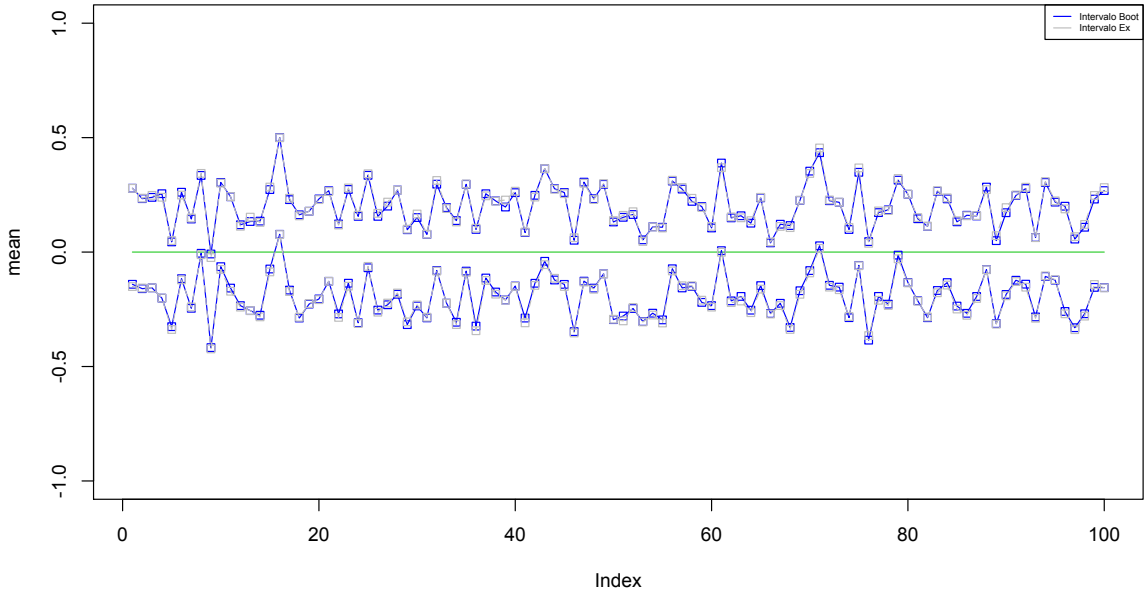

Figura 1: Comparación de intervalos. Usando Boos y (2).

Hay otra manera de crear intervalos de confianza. En la literatura se le conoce como el método BC_a (*bias corrected and accelerated*) y es el siguiente algoritmo.

Algoritmo 3. Para estimar un intervalo de confianza para θ.

1. Dar una m.o. $x_1,\ldots,x_m$ y $\hat{\theta}$ un estimador *plug-in* de θ.
2. Generar, para $j=1,\ldots,N$,
$$x^{*(j)} = \{x_{j_1},\ldots,x_{j_m}\},$$
donde $j_1,\ldots,j_m \sim U\{1,\ldots,m\}$.
3. Hacer, para $j=1,\ldots,N$
$$\hat{\theta}^{*(j)} = \hat{\theta}(x^{*(j)}).$$

4. Hacer $\hat{z} = \Phi^{-1}\left(\frac{\#\{j;\hat{\theta}^{*(j)} \leq \hat{\theta}\}}{N}\right)$
5. Para $i = 1, \ldots, m$, hacer
$$\theta_{(i)} = \hat{\theta}_{n-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n),$$
donde $\hat{\theta}_{n-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ es la evaluación de $\hat{\theta}$ en la muestra pero sin el valor x_i, por eso el subíndice $n-1$.
6. Calcular
$$\bar{\theta}_{(\cdot)} = \frac{1}{m}\sum_{i=1}^{m} \theta_{(i)}.$$
7. Hacer
$$\hat{a} = \frac{1}{6}\frac{\sum_{i=1}^{m}(\bar{\theta}_{(\cdot)} - \theta_{(i)})^3}{(\sum_{i=1}^{m}(\bar{\theta}_{(\cdot)} - \theta_{(i)})^2)^{3/2}},$$
$$q_l = \Phi\left(\hat{z} + \frac{\hat{z} + \Phi^{-1}(\alpha/2)}{1 - \hat{a}(\hat{z} + \Phi^{-1}(\alpha/2))}\right), \text{ y}$$
$$q_u = \Phi\left(\hat{z} + \frac{\hat{z} + \Phi^{-1}(1-\alpha/2)}{1 - \hat{a}(\hat{z} + \Phi^{-1}(1-\alpha/2))}\right).$$
8. Ordenar de menor a mayor la muestra $\{\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(N)}\}$. Digamos que la muestra ordenada es $\hat{\theta}^{*}_{(1)}, \ldots, \hat{\theta}^{*}_{(N)}$.
9. Hacer $l = \lceil q_l N \rceil$ y $u = \lceil q_u N \rceil$.
10. Y obtenemos el intervalo de confianza
$$\left(\hat{\theta}^{*}_{(l)}, \hat{\theta}^{*}_{(u)}\right).$$

Problema 1. *Implemente al algoritmo (3). Suponga que tiene une m.o. N(0,1) y estime un intervalo de confianza para la media. Compare éste intervalo con el intervalo que se da en la expresión (2). Igual que en el ejemplo anterior, haga varias repeticiones y grafique.*