Estimación del error estándar de la media

Usando el principio de "plug-in" podemos reemplazar la empírica $\hat{F}_n$ por F en la fórmula $se_F(\overline{X}) = \sigma_F / \sqrt{n}$. Como el estimado "plug-in" de σ_F está dado por

$$\hat{\sigma} = \sigma_{\hat{F}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$ entonces

$$\hat{se}(\overline{x}) = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n^2}}$$

Notar que este estimador es distinto al estimado usual del error estándar puesto que generalmente σ_F es estimado por

$$\tilde{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Hay muy pocos estimadores que tienen una forma explicita para su error estándar. Generalmente, lo que existen son resultados asintóticos que basan en el hecho de que $\sqrt{n}(\hat{\theta} - \theta) \to N(0, \frac{1}{I(\theta)})$ (Teorema de Mann-Withney), donde I(θ)= -E[$\frac{\partial^2}{\partial \theta^2} \log f(x;\theta)$] es llamada la información de Fisher. Aquí $f(x,\theta)$ representa la función de densidad de probabilidad de X que tiene como parámetro θ.

En palabras mas sencillas Var($\hat{\theta}$)= $\frac{1}{nI(\theta)}$ cuando n es grande.

Ejemplo 12. Cuando n es grande Var(Mediana)= $\frac{1}{4nf^2(\xi_{1/2})}$, donde f($\xi_{1/2}$) representa el valor de la función de densidad en el centro de la distribución. Este hecho resulta como un caso particular de la distribución asintótica de los quantiles de una distribución. (Ver la prueba en Serfling, Approximation theorems of Mathematical Statistics). Dado un valor p entre 0 y 1 un quantile es un valor ξ tal que F(ξ)=p. En particular, cuando p=.5 se obtiene la mediana. Se usa ξ_p para representar el quantil del p%.

Asumiendo que la población es normal, $f(\xi_{1/2}) = \frac{1}{\sigma\sqrt{2\pi}}$. Luego, Var(mediana)= $\frac{\sigma^2 2\pi}{4n}$, y por lo tanto el error estándar de la mediana sería aproximadamente, $\sigma\sqrt{\frac{\pi}{2n}}$. Es obvio, que para hacer este cálculo aproximado hay que hacer bastante manipuleo algebraíco y cotejar si algunas suposiciones se aplican.

Con el método bootstrap se pueden estimar errores estándar de cualquier estadístico $\hat{\theta}$, no necesariamente de la media , no importa lo complicado que sea su funcional.

El estimado bootstrap del error estándar de un estimador.

Supongamos que se ha observado una muestra **x**=(X_1,X_2,…X_n) de una población que tiene función de distribución F ,la cual es desconocida y que se desea estimar un parámetro de interés θ=t(F) basado en la muestra tomada x. Para esto hay que encontrar un estadístico $\hat{\theta} = s(\mathbf{x})$. Lo que desea es determinar cuán preciso es $\hat{\theta}$ mediante la estimación de su error estándar. El Bootstrapping, introducido por Efron en 1979, es un método para estimar el error estándar de un estimador basado enteramente en el uso de computadoras y sin importar la complejidad matemática de $\hat{\theta} = s(\mathbf{x})$. No requiere de cálculos teóricos.

Muestra Bootstrap. Dada la muestra muestra **x**=(X_1,X_2,…X_n) , una muestra bootstrap es una muestra **x***=($X_1^*, X_2^*,...X_n^*$) del mismo tamaño y escogida con reemplazo. Es como si se estuviera tomando muestras con reemplazo y del mismo tamaño de la población **x**=(X_1,X_2,…X_n) que tiene como función de distribución a la distribución Empírica $\hat{F}_n$. Sea $\hat{\theta}^* = s(x^*)$, el valor del estimador en la muestra bootstrap **x***, entonces el estimador B Bootstrap ideal del error estándar del estadístico $\hat{\theta} = s(\mathbf{x})$ estará dado por

$$s\hat{e}(\hat{\theta}) = se_{\hat{F}_n}(\hat{\theta}^*)$$

En otras palabras el estimado bootstrap del estándar error es el error estandar de $\hat{\theta} = s(\mathbf{x})$ para todo el conjunto posible de muestras bootstrap elegidas de la muestra original. Si la muestra original tiene m datos, entonces habría en total n^n muestras que se podrian tener. En la practica uno toma B muestras bootstrap y calcular el error estándar estimado de $\hat{\theta}^* = s(x^*)$ en base a las B muestras bootstrap. Esto es,

$$\hat{se}_B(\hat{\theta}) = \sqrt{\frac{\sum_{b=1}^{B} (\hat{\theta}^*(b) - \bar{\hat{\theta}}^*)^2}{B-1}}$$

donde $\bar{\hat{\theta}}^* = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$.

El limite de $\hat{se}_B(\hat{\theta})$ cuando B tiende a infinito es el estimado bootstrap ideal $se_{\hat{F}}(\hat{\theta}^*)$.

Esto es, $\lim_{B\to\infty} \hat{se}_B = se_{\hat{F}}(\hat{\theta}^*)$.

El estimado bootstrap ideal y su aproximación $\hat{se}_B(\hat{\theta})$ son llamados estimados bootstrap noparamétricos porque están basados en la función de distribución empírica que es un estimador de F. Existen también estimados bootstrap paramétrico que usa otro tipo de estimado para F.

El Algoritmo Bootstrap para estimar errores estándar

1 Seleccionar B muestras bootstrap independientes, $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ cada una consistente de n valores extraídos con reemplazo de la muestra original $\mathbf{x}=(x_1,\ldots x_n)$.

2 Evaluar el estadístico en cada muestra bootstrap. Esto es

$\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$ para b=1,2,…..B.

3. Estimar el error estándar $se_F(\hat{\theta})$ por la desviación estándar de las repeticiones del estadístico en las B muestras bootstrap. Es decir, por

$$\hat{se}_B(\hat{\theta}) = \sqrt{\frac{\sum_{b=1}^{B} (\hat{\theta}^*(b) - \bar{\hat{\theta}}^*)^2}{B-1}}$$

con $\bar{\hat{\theta}}^* = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$

La siguiente función en R estima por bootstrappping el error estándar de la mediana para un número genérico de muestras bootstrap.

```
Bootmedian=function(data,B)
{nobs=length(data)
bootsample=matrix(0,B,nobs)
for (b in 1:B)
{bootsample[b,]=sample(data,nobs,replace=T)}
medianboot=apply(bootsample,1,median)
seboot=sd(medianboot)
return(seboot)
}
```

Ejemplo 13. Estimar por bootstrapping el error estándar de la mediana de la siguiente muestra del nivel de colesterol de 8 personas

233, 259, 215, 322, 289, 220, 276, 299

Comparar el resultado con el resultado asintótico teórico asumiendo que la población es normal.

Solución:

Para distintos valores de muestras bootstrap obtenemos los siguientes resultados

B	50	100	200	500	1000	1500	2000	5000
$\hat{se}_B$	18.46	21.68	20.92	20.47	21.07	20.67	20.92	20.84

B=200 parece ser suficiente para obtener una buena estimación del error estándar

El resultado asintótico estimado será

```
> sd(coles)*sqrt(pi/16)
[1] 17.30889
>
```

Notar que hay cierta diferencia entre los valores estimados de la mediana debido talvez a la alta variabilidad de la muestra original y a que la observación 322 podria ser considerada como “outlier”.

Existen dos librerías en R que llevan a cabo muchos experimentos basados en bootstrapping:

La librería **bootstrap** está basada en el libro "An Introduction to the Bootstrap" by B. Efron and R. Tibshirani, 1993, Chapman and Hall.

La librería **boot** está basada en el material del libro "Bootstrap Methods and Their Applications" by A. C. Davison and D. V. Hinkley (1997, CUP).

La aplicación de bootstrap al ejemplo anterior sería como sigue

```
> library(bootstrap)
bootsample=bootstrap(coles,1000,median)
Warning message:
multi-argument returns are deprecated in: return(thetastar, func.thetastar, jack.boot.val,
jack.boot.se,
> sd(bootsample$thetastar)
[1] 20.02101
>
```

No prestar atención al Warning, porque está librería es algo obsoleta.

La aplicación de boot al ejemplo anterior sería como sigue

```
>library(boot)
>coles.fun=function(data,i)
{d=data[i]
 median(d)}
}
> bootsam1=boot(coles,coles.fun,1000)
> bootsam1

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = coles, statistic = coles.fun, R = 1000)


Bootstrap Statistics :
    original  bias    std. error
t1*    267.5 -3.8235    20.95328
```

Aunque su uso es más indirecto que la función bootstrap, sin embargo da el resultado directamente