Estimación del sesgo por bootstrapping

El Sesgo de un estimador $\hat{\theta}$ es otra medida de precisión.
Sea **x**=(X_1,X_2,…X_n) una muestra aleatoria de una variable aleatoria que tiene distribución F y sea θ=t(F) un parámetro que sea estimar. Se define el sesgo del estimador $\hat{\theta}$=s(**x**) por

$$Sesgo_F(\hat{\theta},\theta)= E_F(\hat{\theta})-\theta=E_F[s(\mathbf{x})]-t(F)$$

Ejemplo: Dada una distribución F , $\hat{\theta}=\bar{x}$ es un estimador insesgado de θ=μ. O sea

$$Sesgo_F(\hat{\theta},\theta)=0$$

Ejemplo: Si F es una distribución uniforme en (0,θ) y $\hat{\theta}=\max(X_1,....,X_n)$ entonces

$$Sesgo_F(\hat{\theta},\theta)= E_F(\hat{\theta})-\theta=\frac{n}{n+1}\theta-\theta=-\frac{1}{n+1}\theta$$

El estimado bootstrap ideal del sesgo se obtiene sustituyendo F por su distribución empírica $\hat{F}_n$ está dado por

$$Sesgo_{\hat{F}_n}(\hat{\theta},\theta)=E_{\hat{F}_n}[s(x^*)]-t(\hat{F}_n)$$

donde **x*** es una muestra bootstrap de **x.** Aquí el valor esperado se aproxima usando todas las muestras con reemplazo que se pueden extraer de la muestra original.
En la practica el estimado bootstrap ideal del sesgo es aproximado tomando B muestras bootstrap de la muestra original **x**. Así el estimado bootstrap de θ basado en B repeticiones está dado por

$$Sesgo_B(\hat{\theta},\theta)=\bar{\hat{\theta}}^*-t(\hat{F}_n)$$

donde $\bar{\hat{\theta}}^*$ es el promedio de las estimaciones de $\hat{\theta}$=s(**x**) en cada muestra bootstrap.

Ejemplo. Dada la siguiente muestra

23 37 19 40 24 9 13 4

estimar por bootstrap el sesgo del estimador del máximo usando 2000 muestras bootstrap.

```
sesgomaxboot=function(muestra,B){
nobs=length(muestra)
bootsample=matrix(0,B,nobs)
for(b in 1:B)
{
```

```
bootsample[b,]=sample(muestra,nobs,replace=T)
}
bootmax=apply(bootsample,1,max)
sesgo=mean(bootmax)-max(muestra)
sesgo
}

x=c(23, 37, 19, 40, 24, 9, 13, 4)
> sesgomaxboot(x,2000)
[1] -2.4415
```

El teórico es -40/9= -4.44444

Ejemplo. Estimador de razón. Los siguientes datos representan la población en miles de 10 ciudades de US según los censos de 1920(x) y 1930(y).

i	1	2	3	4	5	6	7	8	9	10
X	138	93	61	179	48	37	29	23	30	2
Y	143	104	69	260	75	63	50	48	111	50

La población total en 1930 se puede estimar multiplicando la población total en 1920 por el estimador θ=E(Y)/E(X), el cual es llamado un estimador de razón.

Dada un par de variables aleatorias X y Y distribuidas conjuntamente se desea estimar el parámetro $\theta = \frac{E(Y)}{E(X)}$, el estimador muestral será $\hat{\theta} = \frac{\overline{y}}{\overline{x}}$.

La siguiente función en R halla el estimador por bootstrapping del error estándar y sesgo de un estimador de razón usando B muestras bootstrap.

```
 bootratio=function(data,B)
 {# esta funcion halla el error estandar y el sesgo estimado
 #por bootstrapping de una razon
 nobs=dim(data)[1]
 thetaest=mean(data[,2])/mean(data[,1])
 bootrat=rep(0,B)
 for (b in 1:B)
 {bootindex=sample(1:nobs,nobs,replace=T)
 bootrat[b]=mean(data[bootindex,2])/mean(data[bootindex,1])
 }
 seboot=sd(bootrat)
 cat("El error estandar estimado por bootstrapping de la razon es",seboot,"\n")
 biasboot=mean(bootrat)-thetaest
 cat("El sesgo estimado por bootstrapping de la razon es",biasboot,"\n")
 }
```

Aplicando la función a los datos con B=2000 y B=10000 muestras bootstrap se obtiene

> brazon=bootratio(pobla,2000)
El error estandar estimado por bootstrapping de la razon es 0.2264477
El sesgo estimado por bootstrapping de la razon es 0.04603658
La razón del sesgo con respecto al error estándar es
> .0460/.2264
[1] 0.2031802

brazon=bootratio(pobla,10000)
El error estandar estimado por bootstrapping de la razon es 0.2191757
El sesgo estimado por bootstrapping de la razon es 0.03769554

La razón del sesgo con respecto al error estándar es

> .0376/.2191
[1] 0.1716111

Usualmente el número de muestra bootstrap necesarias para estimar el sesgo es mayor que el número de muestras requeridas para estimar el error estándar, debido es la mayor variablidad presente en la estimación del sesgo como lo muestra los siguientes resultados para el ejemplo anterior.

B	100	500	1000	2000	5000
CV(se_B)	0.0920	0.0550	0.0438	0.0306	0.0260
CV($Sesgo_B$)	0.3826	0.2782	0.1502	0.1175	0.0973

Ejemplo: Estimación por Boostrapping del error estándar y sesgo del coeficiente de asimetría (skewness)

Sea X una variable aleatoria con media μ entonces la medida de asimetría de su distribución se define por

$$\gamma = \frac{E(X-\mu)^3}{[E(X-\mu)^2]^{3/2}}.$$

Usando el método de estimación por momentos para el segundo y tercer momento central se tiene que la asimetría puede ser estimada por

$$skew = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}}{\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}\right]^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

Hay muchas otras variantes de la formula. Recordando que $\sigma^2 = E(X - \mu)^2$ y que *s* es un estimado de σ,produce el siguiente estimado

$$skew = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$$

Esta forma de estimado es el que usa la función skewness de la librería fbasics de R.
La siguiente función en R calcula el skewness de una variable aleatoria X.

```
skewness=function (x)
{a=sqrt(length(x))*sum((x-mean(x))^3)
b=(sum((x-mean(x))^2))^1.5
skew=a/b
}
```

La siguiente función de R calcula la estimación por boostrapping del error estándar y del sesgo del skewness de una variable aleatoria X.

```
bootskew=function(data,B)
{# esta funcion halla el error estandar y el sesgo estimado
#por bootstrapping de una razon
nobs=length(data)
thetaest=skewness(data)
bootskew=rep(0,B)
for (b in 1:B)
{bootsample=sample(data,nobs,replace=T)
bootskew[b]=skewness(bootsample)
}
seboot=sd(bootskew)
cat("El error estandar estimado por bootstrapping del skewness es",seboot)
biasboot=mean(bootskew)-thetaest
cat("\nEl sesgo estimado por bootstrapping del skewness es",biasboot,"\n")
}
```

Una aplicación a las dos variables del conjunto brain produce los siguientes resultados

```
bootskew(brain[,1],100)
El error estandar estimado por bootstrapping del skewness es 0.2704018
El sesgo estimado por bootstrapping del skewness es -0.01824653
```

> bootskew(brain[,1],500)
El error estandar estimado por bootstrapping del skewness es 0.2852596
El sesgo estimado por bootstrapping del skewness es 0.01726378
> bootskew(brain[,1],1000)
El error estandar estimado por bootstrapping del skewness es 0.2995121
El sesgo estimado por bootstrapping del skewness es -0.009411089

> bootskew(brain[,2],100)
El error estandar estimado por bootstrapping del skewness es 0.2770318
El sesgo estimado por bootstrapping del skewness es -0.04094157
> bootskew(brain[,2],500)
El error estandar estimado por bootstrapping del skewness es 0.2438393
El sesgo estimado por bootstrapping del skewness es -0.03295653
> bootskew(brain[,2],1000)
El error estandar estimado por bootstrapping del skewness es 0.2566760
El sesgo estimado por bootstrapping del skewness es -0.04754999
>

Solamente en el caso de estimadores "plug-in" existe una mejor manera de estimar el sesgo por bootstrapping usando un concepto llamado de "vector de remuestreo" pero no será discutida aquí (ver Efron y Tibshirani, pags 130-133).

Estimado bootstrap de un estimador corregido por sesgo

La razón principal de estimar el sesgo es corregir el estimador inicial $\hat{\theta}$ haciéndolo menos sesgado. Luego el estimador de θ corregido por sesgo será

$$\bar{\theta} = \hat{\theta} - Sesgo_B(\hat{\theta}, \theta)$$

O equivalentemente,

$$\bar{\theta} = \hat{\theta} - (\bar{\hat{\theta}}^* - \hat{\theta}) = 2\hat{\theta} - \bar{\hat{\theta}}^*$$

En el ejemplo de la razón de poblaciones se tiene que $\hat{\theta} = 1.5203$ y el estimado del sesgo basado en 2000 muestras bootstrap es .0460. Luego el estimador corregido por sesgo será

$$\bar{\theta} = 1.5203 - .0460 = 1.4743$$