Intervalos de confianza basado en bootstrap

Consideremos una variable aleatoria X que tiene una función de distribución $F_\theta(x)=F(x-\theta)$, donde θ es llamado un parámetro de localización. Ejemplos de parámetros de localización son la media y la mediana.
Una forma básica de predecir el comportamiento de θ es usando un intervalo de la forma

$$(\hat{\theta}\text{-Error},\ \hat{\theta}\text{+Error})$$

donde el término Error depende de la distribución del estimador y del error estándar del mismo. El intervalo de confianza siempre va asociado a un nivel de confianza $100(1-\alpha)\%$ con que se espera caiga el parámetro en el intervalo. Aquí α es llamado el nivel de significación que varía entre 0 y 1, siendo los valores más usados .10,.05 y .01, los cuales corresponden al 90, 95 y 99% de confianza respectivamente.

Así un intervalo de confianza para la media poblacional μ es de la forma

a) $$(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

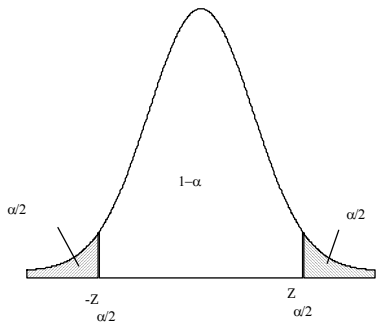
si la población es normal o si la muestra es grande con varianza poblacional conocida

b) $$(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}})$$

si la población es normal o si la muestra es grande con varianza poblacional desconocida

c) $$(\bar{x} - t_{(\alpha/2,n-1)})\frac{s}{\sqrt{n}}, \bar{x} + t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}})$$

si la población es normal, la muestra es pequeña y la varianza poblacional desconocida.
Los valores $z_{\alpha/2}$ y $t_{\alpha/2}$ están asociados con los percentiles de la normal estándar Z y de t respectivamente. Así $z_{\alpha/2}$ representa el percentile del $100(1-\alpha/2)\%$ de la normal estándar como se muestra en la siguiente figura.



Ejemplo: Para los datos del colesterol de pacientes hallar un intervalo de confianza del 95% para la media poblacional.

Solución: Usando la función **t.test** de R se obtiene el siguiente resultado

pruebat=t.test(coles)

> pruebat$conf
[1] 231.4684 296.7816
attr(,"conf.level")
[1] 0.95

Es decir, hay una confianza del 95% de que la media poblacional caiga entre 231.4 y 296.7.

Intervalos de confianza por bootstrapping: Método estándar

Es el método más básico de estimar intervalos de confianza para un parámetro de localización. θ. Asumiendo que la distribución del estadístico $\hat{\theta}$ es aproximadamente normal se tiene que el intervalo de confianza del 100(1-α)% para θ es de la forma

$$(\hat{\theta} - z_{\alpha/2} s.e(\hat{\theta}), \hat{\theta} + z_{\alpha/2} s.e(\hat{\theta}))$$

La forma estándar de estimar un intervalo de confianza usando bootstrapping es simplemente reemplazando en la fórmula anterior el error estándar por su estimado bootstrap.
La siguiente función en R encuentra un intervalo de confianza usando el metodo estándar de bootstrapping

```
bootmean.ic.std=function (data,level,B)
{alpha=1-.01*level
lq=qnorm(alpha/2)
uq=qnorm(1-.5*alpha)
seb=bootmean(data,B)
l.ic=mean(data)+lq*seb
u.ic=mean(data)+uq*seb
cat("El intervalo de confianza del",level,"% para la media es:(",l.ic,",",u.ic,")\n")
}
```

Ejemplo: Para los datos de colesterol calcular un intervalo de confianza para hallar un intervalo de confianza del 95% por el método estándar con 1000 muestras bootstrap

Solución: Usando la función **bootmean.ic.std** se obtiene el siguiente resultado
bootmean.ic.std(coles,95,1000)
El intervalo de confianza del 95 % para la media es:(238.501 , 289.749)
>

Es decir, hay una confianza del 95% de que la media poblacional caiga entre 238.50 y 289.74

Cálculo de intervalo de confianza usando Bootstrapping: Método de los percentiles

La idea aquí es estimar la función de distribución F del estimador $\hat{\theta}$ usando Bootstrapping. Luego, el intervalo de confianza del 100(1-α) para θ será simplemente

$$[\hat{F}^{-1}(\alpha/2),\ \hat{F}^{-1}(1-\alpha/2)]$$

donde $\hat{F}^{-1}(\alpha/2)$ representa el percentil del α/2 de la distribución del estimador por bootstrapping $\hat{\theta}*$, es decir un valor tal que la probabilidad acumulada hasta dicho valor sea α/2, y $\hat{F}^{-1}(1-\alpha/2)$ representa el percentil del 1-α/2.
El procedimiento consiste en generar primero una cantidad B (B aprox 1000) de muestras bootsrap. Luego se calcula en cada una de ellas el estadístico que estima al parámetro y finalmente se calculan los percentiles de la distribución basada en B valores como se ha hecho enlos anteriores casos

Ejemplo: Hallar el intervalo de confianza para la muestra de colesterol usando Bootstrapping con el método de los percentiles.

Solución: Hemos construido una función **bootmean.ic.perc** en R que resuleve el problema

```
bootmean.ic.perc=function(data,nboot,level)
{
# Esta funcion calcula el intervalo de confianza
# para cualquier estimado usando Bootstrapping con
# el método de los percentiles
# Level representa el nivel de confianza y esta dado en
# porcentaje
nobs=length(data)
alpha=1-.01*level
bootsample=matrix(0,nboot,nobs)
for(i in 1:nboot)
{bootsample[i,]=sample(data,nobs,replace=T)}
meanboot=apply(bootsample,1,mean)
meanboot=sort(meanboot)
kperc=floor((nboot+1)*alpha*.5)
l.ic=meanboot[kperc]
u.ic=meanboot[nboot+1-kperc]
cat("El intervalo de confianza del",level,"% para la media
es:(",l.ic,",",u.ic,")\n")
}
```

Aplicando la función con 1000 muestras bootstrapeada y un nivel de confianza del 95% se obtiene

```
bootmean.ic.perc(coles,1000,95)
El intervalo de confianza del 95 % para la media es:( 240.625 , 291.125 )
>
```

Así que el intervalo de confianza aproximado saldrá (240,625, 291.125), bastante cerca del que dice la teoría estadistica.

Ejemplo. Consideremos la siguiente muestra proveniente de una distribución exponencial con parámetro 1.

2.3795 0.1912 0.3136 1.9891 0.3222 0.6468 0.2039 2.5263 0.2780 1.1861 0.1949
1.2198 2.4566 0.3028 0.5941

Estimar por Boostrapping un intervalo de confianza del 95% para la media poblacional θ.

Solución:
> bootmean.ic.perc(x,1000,95)
El intervalo de confianza del 95 % para la media es:(0.5594267 , 1.448447)
>

Intervalos de confianza por Bootstrapping estudentizado o Bootstrapping-t.

La idea aqui es sustituir los percentiles de la distribución t por los percentiles de la distribución de los valores estudentizados de estadístico $\hat{\theta}$ en la muestras bootstrap. Es decir de los valores definidos por

$$t^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\hat{se}(\hat{\theta}^*(b))}$$

donde $\hat{\theta}^*$(b) y $\hat{se}(\hat{\theta}^*$(b)) son los valores del estimador y valor estimado del error del estimador en la b-ésima muestra bootstrap respectivamente; $\hat{\theta}$ es el estimado en la muestra original. Si se está tratando de estimar la media el error estándar estimado sería $\frac{s(\hat{\theta}^*)}{\sqrt{n}}$, pero en el caso de otros estimadores tales como la mediana habría que aplicar otro bootstrapping para estimar este error estándar.

Luego, el Intervalo de confianza del 100(1-α)% por bootstrapping estudentizado estará dado por

$$(\hat{\theta} + \hat{F}_t^{-1}(\alpha/2)^* \hat{se},\ \hat{\theta} + \hat{F}_t^{-1}(1\text{-}\alpha/2)^* \hat{se})$$

donde $\hat{F}_t^{-1}$ (α/2) representa el percentil del α/2 de la distribución de t*, es decir un valor tal que la probabilidad acumulada hasta dicho valor sea α/2, y $\hat{F}_t^{-1}$ (1-α/2) representa el percentil del 1-α/2.
Mas precisamente, si k=[(B+1)α/2] representa el entero más grande menor o igual que (B+1) α/2. Entonces

$\hat{F}_t^{-1}(\alpha/2)$ = k-ésima observación más grande de t*(b), y
$\hat{F}_t^{-1}(1-\alpha/2)$ = (B+1-k)-ésima observación más grande de t*(b)

Aqui, $\hat{se}$ representa el error estándar del estimador de la muestra tomada, nuevamente si es la media no es problema, pero si es otro estimador habría que aplicar bootstrapping o Jackknife (a ser visto más adelante) para estimarlo.

La siguiente función bootmean.ic en R calcula intervalos de confianza para la media

```
bootmean.ic.stu=function (data,conf,B)
{# Esta funcion calcula intervalo de confianza para la media
 # por le metodo de bootstrapping estudentizado
# data: nombre del conjunto de datos
#conf= El nivel de confianza,por ejemplo 95
#B : el numero de muestras bootstrap
nobs=length(data)
alpha=1-.01*conf
bootsample=matrix(0,B,nobs)
for(i in 1:B)
{bootsample[i,]=sample(data,nobs,replace=T)}
meanboot=apply(bootsample,1,mean)
seboot=apply(bootsample,1,sd)/sqrt(nobs)
pivotb=(meanboot-mean(data))/seboot
kperc=floor((B+1)*alpha*.5)
pivotb=sort(pivotb)
lq=pivotb[kperc]
uq=pivotb[B+1-kperc]
# Las 4 lineas anteriores pueden ser sustituidas  por las siguientes dos y dan una buena
#aproximacion
#lq=quantile(pivotb,alpha/2)
#uq=quantile(pivotb,1-(alpha)/2)
l.ic=mean(data)+lq*sd(data)/sqrt(nobs)
u.ic=mean(data)+uq*sd(data)/sqrt(nobs)
cat("El intervalo de confianza del",conf,"% para la media es:(",l.ic,",",u.ic,")\n")}
```

Ejemplo: Para los datos de colesterol,calcular un intervalo de confianza para hallar un intervalo de confianza estudentizado del 95% usando 1000 muestras bootstrap

Solución: Usando la función **bootmean.ic.stu** se obtiene el siguiente resultado

```
> bootmean.ic.stu(coles,95,1000)
El intervalo de confianza del 95 % para la media es:( 232.1805 , 302.8118 )
>
```

Es decir, hay una confianza del 95% de que la media poblacional caiga entre 232.18 y 302.81

Intervalos de confianza por bootstrapping estudentizado con estabilización de la varianza

Tibshirani (1988) notó que algunas veces usar boostrapping estudentizado podría producir intervalos de confianza ineficientes para un parámetro θ, como por ejemplo intervalos de confianza más grandes que (-1,1) para los coeficientes de correlación. El mostró que una mejor alternativa es transformar el parámetro θ , usando una función g. Luego, una vez hallado el

intervalo de confianza para g(θ) se toma la inversa y se obtiene el intervalo de confianza pedido para θ. La transformación g ideal es aquella que estabiliza la varianza en palabras sencillas es aquella tal que la varianza de g($\hat{\theta}$) sea aproximadamente constante.

Propiedad. Sea X una variable aleatoria con media θ y desviación estándar s(θ) y sea g una función continua y derivable entonces

$$Var(g(X))\approx[g'(\theta)]^2 Var(X)$$

Luego, si se define g(x)=$\int^{x} \frac{1}{s(u)}du$. Se obtendría que Var(g(X))≈constante. El problema está que algunas veces la relación entre la media y la desviación estandar no viene dada en forma explicita (inclusive en el caso paramétrico) y habria que estimarla también, en forma similar a como se estima un modelo de regresión.

Finalmente, el procedimiento para hallar Intervalos de confianza, usando Bootstrapping Estudentizado con estabilización de varianza seria como sigue:

1. Extraer B_1 (B_1=100 es suficiente) muestras bootstrap de la muestra original y en cada una de ellas estimar θ por $\hat{\theta}$*.
2. Tomar B_2 (B_2=25 es suficiente) muestras bootstrap de cada una de las muestras bootstrap obtenidas en el paso 1 y estimar el error estándar por se($\hat{\theta}$*).
3. Ajustar una curva a los puntos ($\hat{\theta}$*,se($\hat{\theta}$*)) para producir un estimado de la función se(θ). Aqui se puede usar regresión lineal, polinómica, o cualquier regresión noparamétrica tal como loess, splines, ksmooth, etc.
4. Estimar la transformación que estabiliza la varianza, g(θ) usando la definición anterior e integración numérica ya que es bien probable que g sea estimada con una tabla de valores y no en forma explicita.
5. Usando B_3 (B_3=1000 es suficiente) nuevas muestras bootstrap, calcular un intervalo de confianza para g(θ) pero considerando el pivote t=g($\hat{\theta}$*)-g($\hat{\theta}$), puesto que el error estándar que es constante se puede considerar como 1.
6. Finalmente, aplicar g^{-1} a cada extremo del intervalo anterior para encontrar el intervalo de confianza para el parámetro θ basado en la muestra tomada.

Ejemplo. Consideremos la muestra proveniente de una distribución exponencial con parámetro 1, usada en un ejemplo anterior. Estimar por Boostrapping un intervalo de confianza del 95% para la media poblacional θ.

Solución:

Es conocido que si X_1,….X_n es una muestra de una variable exponencial con parámetro θ entonces X_1+…..X_n se distribuye como una Gamma(n,θ), luego, si

$\hat{\theta} = \overline{X}$, entonces $E(\hat{\theta}) = \theta$ y $se(\hat{\theta}) = \frac{\theta}{\sqrt{n}}$. Luego la transformación que estabilizaría la varianza sería $g(\theta) = \int^{\theta} \frac{\sqrt{n}}{u} du = \sqrt{n}\log(\theta)$.

Un intervalo de confianza para θ será de la forma

$(\hat{\theta} - t_{(\alpha/2)} se(\hat{\theta}), \hat{\theta} + t_{(\alpha/2)} se(\hat{\theta}))$

el cual será estimado a través de

$$(g(\hat{\theta}) + \hat{F}^{-1}(\alpha/2) se(g(\hat{\theta}^*)), \hat{\theta} + \hat{F}^{-1}(1-\alpha/2) e(g(\hat{\theta}^*)))$$

Finalmente se aplica g^{-1} a cada extremo del intervalo de confianza para obtener el intervalo de confianza para θ.

La siguiente función en R encuentra el intervalo de confianza para θ usando estabilización de la varianza.

```
bootmean.ic.vs=function(data,nboot1,nboot2,nboot3,level)
{# Esta funcion calcula el intervalo de confianza
# para cualquier estimado usando Bootstrapping estudentizado
# con estabilización de la varianza
# data: nombre delconjunto de datos
# nboot1: número de muestras bootstrap para estimar theta
# nboot2: número de muestras bootstrap para estimar se(theta)
# nboot3: número de muestras bootstrap para estimar la distribución del pivote
# level: representa el nivel de confianza

nobs=length(data)
alpha=1-.01*level
bootsample=matrix(0,nboot1,nobs)
seb=rep(0,nboot1)
for(i in 1:nboot1){
bootsample[i,]=sample(data,nobs,replace=T)
seb[i]=bootmean(bootsample[i,],nboot2)
}
bmean=apply(bootsample,1,mean)
stheta=bmean/sqrt(nobs);
ghat=sqrt(nobs)*log(mean(data))
par(mfrow=c(1,2))
plot(bmean,seb,main=expression(s(theta)))
lines(lowess(bmean,seb))
lines(bmean,stheta)
bmeans=sort(bmean)
gtheta=sqrt(nobs)*log(bmeans);
plot(bmeans,gtheta,main=expression(g(theta)),xlab="bmean",type="l");
bootsample1=matrix(0,nboot3,nobs)
for(i in 1:nboot3){
```

```
bootsample1[i,]=sample(data,nobs,replace=T)
}
bmean1=apply(bootsample1,1,mean)
pivotb=sqrt(nobs)*log(bmean1)-ghat
pivotb=sort(pivotb)
kperc=floor((nboot3+1)*alpha*.5)
lq=pivotb[kperc]
uq=pivotb[nboot3+1-kperc]
l.ic=ghat+lq*sd(sqrt(nobs)*log(bmean1))
u.ic=ghat+uq*sd(sqrt(nobs)*log(bmean1))
l.ic=exp(l.ic/sqrt(nobs))
u.ic=exp(u.ic/sqrt(nobs))
cat("El intervalo de confianza del",level,"% para la media es:(",l.ic,",",u.ic,")\n")
}
```

Se obtienen los siguientes resultados

```
> bootmean.ic.vs(x,100,20,1000,95)
El intervalo de confianza del 95 % para la media es:( 0.6146817 , 1.380337 )
>
```

Mientras que si se usa el metodo de doble bootstrapping o boostraping anidado, es decir aquel donde se usa una muestra bootstrap para estimar θ y una segunda muestra bootstrap para estimar el error estándar de θ, se obtiene el intervalo

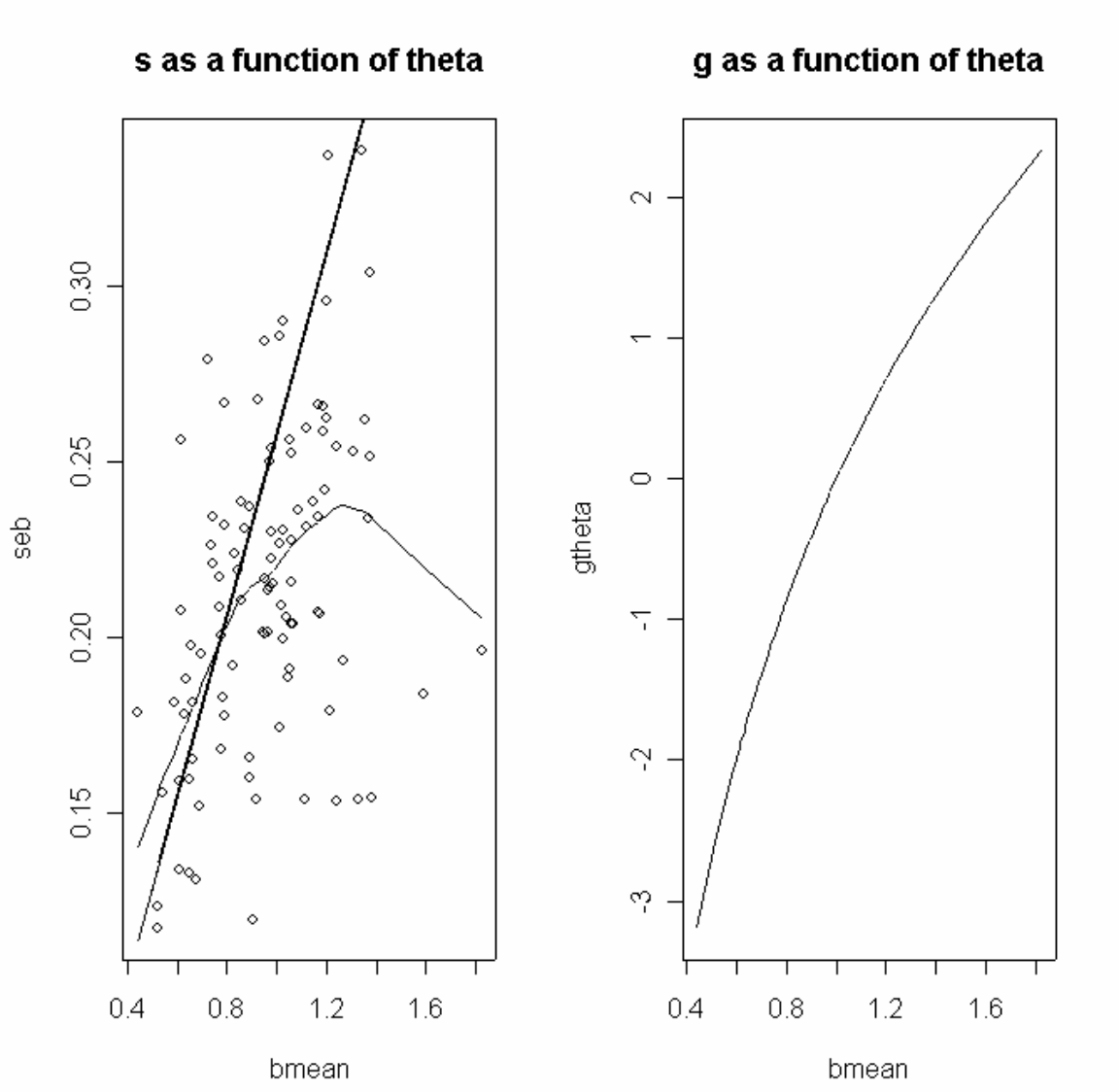
```
bootmean.ic.db=function(data,nboot1,nboot2,level)
{
#Esta funcion calcula el intervalo de confianza
#para un parametro usando Bootstrapping doble (o anidado)
#level representa el nivel de confianza
#Llama a la funcion bootmean que estima el error estandar de
# la media por bootstrapping

nobs=length(data)
alpha=1-.01*level
bootsample=matrix(0,nboot1,nobs)
seb=rep(0,nboot1)
for(i in 1:nboot1)
{bootsample[i,]=sample(data,nobs,replace=T)
seb[i]=bootmean(bootsample[i,],nboot2)
}
meanboot=apply(bootsample,1,mean)
pivotb=(meanboot-mean(data))/seb
kperc=floor((nboot1+1)*alpha*.5)
pivotb=sort(pivotb)
lq=pivotb[kperc]
uq=pivotb[nboot1+1-kperc]
serr=sd(meanboot)
l.ic=mean(data)+lq*serr
u.ic=mean(data)+uq*serr
cat("El intervalo de confianza del",level,"% para la media
es:(",l.ic,",",u.ic,")\n")
```

}

bootmean.ic.db(x,1000,25,95)
El intervalo de confianza del 95 % para la media es:(0.1459485 , 1.483071)
>

el cual es mas amplio que el anterior.

La siguiente figura muestra las relaciones entre $\hat{\theta}*$, se($\hat{\theta}*$) , se(θ) y g(θ)



Existen otros mejores metodos para calcular intervalos de confianza usando boostrapping: el método BC (intervalo de confianza corregido por sesgo y acelerado) y el metodo ABC (intervalo de confianza aproximado por bootstrapping).