Capitulo 3. Teoría y Aplicaciones de Bootstrapping

Referencias:

Efron, B.y Thisbhirani (1993) An introduction to the Bootstrap. Chapman and Hall
Davison, A.C. y Hinkley, D.V. (1997) Bootstrap methods and their applications. Cambridge University Press.

Motivación

Talvez la parte más importante de la estadística es usar los datos de la muestra que se ha recolectado para sacar conclusiones acerca de la población de donde procede la muestra. Por ejemplo, supongamos que a 8 personas que acaba de tener un ataque cardiaco se le tomado una muestra de su colesterol y se obtuvo los siguientes resultados

233 259 215 322 289 220 276 299

El promedio o media de la muestra resulta ser 264.125. Pero uno está interesado realmente en caracterizar el colesterol del total de personas que tiene ataques cardiacos.
Las cantidades que se usan para caracterizar una población son llamados parámetros y se representan por θ. Las cantidades que se calculan usando la muestra tomada se llaman estimados y se representan por $\hat{\theta}$. Se espera que el estimado refleje el comportamiento del parámetro Propiamente $\hat{\theta}$=T(X_1,……X_n) donde (X_1,…..,X_n) representa la muestra aleatoria y T es una función optima que se usa para estimar θ y es llamado el estimador. Notar que el valor del estimado varía con la muestra tomada. Métodos para hallar T son tratados en cursos de Estadística Matemática. Si el parámetro de interés es la media poblacional entonces el estimador óptimo a usar es la media muestral $\bar{x}$. El proceso clásico de estimación requiere asumir una forma distribucional, generalmente Normal o Gaussiana para la población. Sin embargo hoy en día la orientación es usar la data también para estimar esta forma distribucional en lugar de asumirla. Esto es llamado Estadística No-paramétrica.

Muestras aleatorias

Consideremos una población de N unidades U_1,U_2,…U_N, cada una de las cuales es igualmente probable de ser seleccionada en cualquier una extracción al azar que se haga. Una Muestra aleatoria de tamaño n es una colección de n unidades u_1,u_2,…u_n seleccionadas al azar de la población.

Básicamente lo que se selecciona son n enteros j_1,j_2,…j_n entre 1 al N con igual probabilidad 1/N, estos enteros definen los índices de las unidades seleccionadas en la muestra. En principio los enteros j_1,j_2,…j_n se pueden repetir y se dice que la muestra es con reemplazo. Si se desea que sean todos distintos entonces la muestra se dice que es sin reemplazo. Sin embargo cuando el tamaño de muestra es bien pequeño comparado con la población hay una probabilidad muy baja de que haya elementos repetidos en la muestra. En este curso a menos que se diga lo contrario las muestras que se usan son con reemplazo.

Sea x_i las mediciones de interés para la unidad u_i en la muestra. Sea $\mathbf{x}=(x_1,x_2,\ldots,x_n)$ las mediciones observadas en toda la muestra. Por otro lado, sea $\mathbf{X}=(X_1,X_2,\ldots X_N)$ el conjunto de mediciones de todas las unidades de la población o simplemente la población. Entonces, **x** será llamada una muestra aleatoria de **X**.

Ejemplo 1. Extracción de muestras con reemplazo y sin reemplazo usando R

```
> # Definiendo la poblacion
>x=c(9,12,23,45,17,28,15,33)
># muestra  de tamaño 8 con reemplazo
> sample(x,8,replace=T)
[1] 17  9 45  9 45  9 23  9
>#muestra de tamaño 8 sin reemplazo
> sample(x,8,replace=F)
[1]  9 28 45 33 12 17 15 23
```

Inferencia estadistica. Tiene por objetivo determinar que es lo que se puede aprender de la población X en base a la muestra x tomada de ella (extracción de información).
La teoría de probabilidad va en dirección opuesta a inferencia estadística, puesto que a partir de la composición de la población **X** trata de deducir las propiedades de la muestra aleatoria **x** y de funciones T(**x**) obtenidas a partir de ellas. Estas funciones son llamadas estadísticos. Por ejemplo, $T(\mathbf{x})=\bar{\mathbf{x}}=\dfrac{\sum_{i=1}^{n} x_i}{n}$, y $T(\mathbf{x})=s^2=\dfrac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}$.

Variable Aleatoria: es una función que asume sus valores de acuerdo a los resultados de un experimento aleatorio, es decir un experimento en donde hay incertidumbre acerca del resultado que ocurrirá al ejecutarse el experimento.
Una variable aleatoria puede ser discreta o continua de acuerdo a su rango de valores, el cual es llamado espacio muestral y se representa por S.

Ejemplo 2: El mayor de los números que aparece al lanzar al mismo tiempo tres dados legales.
En este caso S={1,2,3,4,5,6}.

Variables aleatorias discretas.
Sea X una variable aleatoria discreta que asume valores $x_1,x_2, \ldots x_n$ y sean las frecuencias f_k=Prob(X= x_k) para k=1,2,….n, entonces la distribución de probabilidades de X está data por

$$\{f_1,f_2,\ldots..f_n\}$$

Ejemplo 3: En el ejemplo 2 las frecuencias están dadas por $f_k=[k^3-(k-1)^3]/6^3$ para k=1,2,…6 y la distribución de probabilidades será $\{f_1,f_2,\ldots f_6\}$. La distribución de

probabilidades puede ser calculada aproximadamente usando simulación. En R usamos la siguiente secuencia de comandos para hacer 1000 simulaciones del experimento

```
> x1=sample(1:6,1000,replace=T)
> x2=sample(1:6,1000,replace=T)
> x3=sample(1:6,1000,replace=T)
> dados=cbind(x1,x2,x3)
> maximos=apply(dados,1,max)
> distest=table(maximos)/1000
> # distribución estimada
> distest
maximos
   1    2    3    4    5    6
0.004 0.043 0.079 0.164 0.285 0.425
>#valores exactos
>trueprob=rep(0,6)
> for(k in 1:6){trueprob[k]=(k^3-(k-1)^3)/216}
> trueprob
[1] 0.00462963 0.03240741 0.08796296 0.17129630 0.28240741 0.42129630
>
```

Ejemplo 4: La distribución binomial con n repeticiones y probabilidad de éxito p tiene frecuencias dadas por

$$f_k = \binom{n}{k} p^k (1-p)^{n-k}$$

para k=0,1,2,….n. y la distribución probabilidades binomiales estará dada por

$$\{f_o, f_2, \ldots . f_n\}$$

Si una variable aleatoria X sigue una distribución binomial se acostumbra a escribir X~Bin(n,p).

Si A es un conjunto de enteros, entonces la probabilidad de que X tome un valor en A será

$$\text{Prob}(X \in A) = P(A) = \sum_{k \in A} f_k$$

Propiedades de f_k:

a) $f_k \geq 0$
b) Si S es el espacio muestral de x entonces, $\sum_{k:x_k \in S} f_k = 1$

Variables aleatorias continuas:

En este caso el espacio muestral S de la variable aleatoria X es un conjunto infinito no enumerable y existe una función nonegativa f, llamada la función de densidad de X tal que

$$P(\boldsymbol{X} \in A) = \int_A f(x)dx \quad \text{para } A \subseteq S.$$

La mas conocidas de todas las distribuciones continuas es la distribución Gaussiana o Normal cuya función de densidad está dada por

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

donde $-\infty<\mu<\infty$, $\sigma>0$ y $-\infty<x<\infty$.

Función de distribución de una variable aleatoria. Sea X una variable aleatoria entonces su función de distribución F está dada por

$$F(x) = P(X \leq x) \quad \text{para } x \in \mathrm{R}.$$

Consideremos una v.a. X con función de distribución F_X, entonces $x_1,\ldots x_n$ será llamada una muestra aleatoria de F_X si para todo i, $F_{xi}=F_X$ y las x_i son independientes entre sí.

Valor Esperado y Varianza de una variable aleatoria

Sea X una v.a con función de distribución F entonces su valor esperado se define por

$$E(X) = \mu = \int_S x dF(x)$$

y su varianza por

$$VAR(X) = \sigma_X^2 = \int_S (x-\mu)^2 dF(x)$$

Es fácil ver que $VAR(X)=E(X^2)-[E(X)]^2$.

En el caso de una variable aleatoria discreta las formulas anteriores se reducen a:

$$E(X) = \sum_{k:x_k \in S} x_k f_k$$

y

$$\sigma_X^2 = \sum_{k:x_k \in S} (x_k - \mu)^2 f_k$$

respectivamente

En el caso de una variable aleatoria continua las formulas anteriores se reducen a:

$$E(x) = \int x f(x) dx$$

y

$$\sigma_X^2 = \int_S (x - \mu)^2 f(x) dx$$

respectivamente.

Ejemplo 5: Si X~Bin(n,p) entonces E(X)=μ=np y $\sigma_X^2 = np(1-p)$.

Ejemplo 6: SI X~N(μ,σ^2) entonces E(X)=μ y $\sigma_X^2 = \sigma^2$.

Ejemplo 7. Sea la v.a. X= 1 si x∈ A⊆S y X=0 en otro caso. Es decir, $X=I_A$, donde I representa a una función indicadora entonces

a) $E(I_A)=P(A)$
b) $VAR(I_A)=P(A)(1-P(A)$

Prueba: Consideremos que la v.a. X es continua entonces

a) $E(I_A) = \int_S I_A(x) f(x) dx = \int_A f(x) dx = P(A)$

b)

$$VAR(I_A) = E(I_A^2) - [E(I_A)]^2 = \int_S I_A^2(x) f(x) dx - P^2(A) = \int_A f(x) dx - P^2(A) = P(A) - P^2(A)$$

Función de distribución Empírica.

La función de distribución (acumulativa) Empírica para una muestra aleatoria $X_1,\ldots X_n$ de F_X está dada por

$$\hat{F}_n(t) = \frac{\#\{i, \quad 1 \le i \le n : X_i \le t\}}{n} = \frac{\#\{X_i \le t\}}{n}$$

Para cada t, $\hat{F}_n(t)$ es un estadístico que da la frecuencia relativa de los valores en la muestra que son menores o iguales que t.

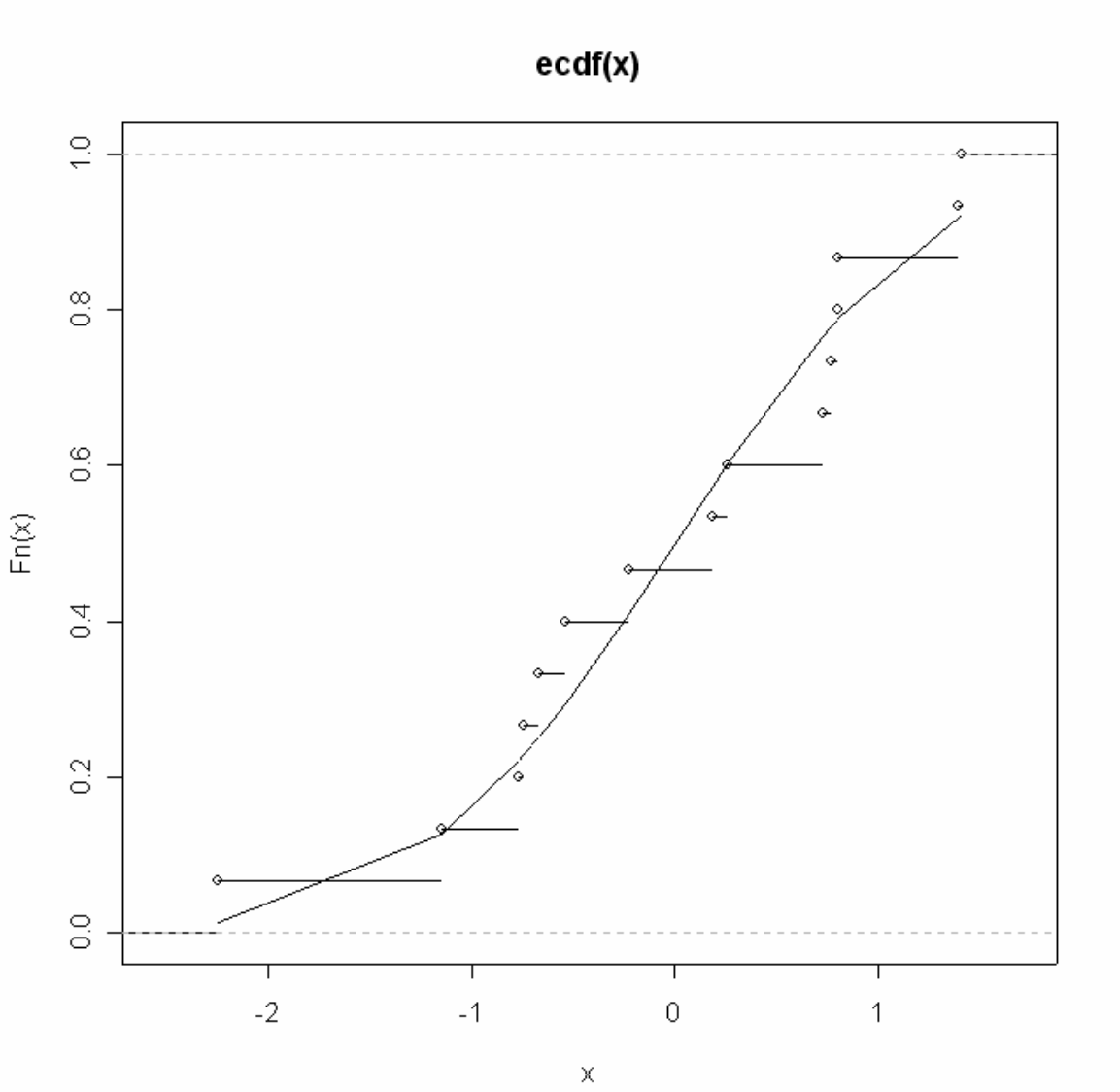
La función de distribución empírica sirve como un estimador noparamétrico de la función de distribución de una variable aleatoria. La definición puede extenderse al caso multivariado.

Ejemplo 8. Comparar gráficamente la verdadera distribución de la curva normal estándar con la distribución empírica usando una muestra aleatoria de tamaño 15.

Primero generamos la muestra aleatoria usando la función **rnorm** y luego usamos la función **ecdf** de la librería **stepfun** en R.

```
library(stepfun)
z=rnorm(15)
plot.ecdf(z)
points(sort(z),pnorm(sort(z)),type="l")
```

La gráfica aparece en la siguiente figura



Notar que la función de distribución Empírica puede escribirse como

$$\hat{F}_n(t) = \frac{\sum_{i=1}^{n} I_{\{Xi \le t\}}(X_i)}{n}$$

Propiedades de la Función de distribución Empírica:

a. $\hat{F}_n$ es creciente de 0 hasta 1.
b. $\hat{F}_n$ es una función escalonada con saltos en los distintos valores de $X_1, X_2, ..., X_n$.
c. $E[\hat{F}_n(t)] = F(t)$.
d. $\text{var}[\hat{F}_n(t)] = F(t)\,[1 - F(t)] / n$.

e. $F_n(x) \rightarrow F(x)$ as $n \rightarrow \infty$ (con probabilidad 1). (Ley fuerte de los grandes números).

Propiedad: Sea g cualquier función de valor real y $\hat{F}_n$ la función de distribución empírica basada en la muestra $x_1, x_2, \ldots x_n$. Entonces

$$\int g(x) d\hat{F}_n(x) = \frac{\sum_{i=1}^{n} g(x_i)}{n}$$