Bootstrap

Georgina Flesia

FaMAF

10 de mayo, 2016

Técnica de bootstrap

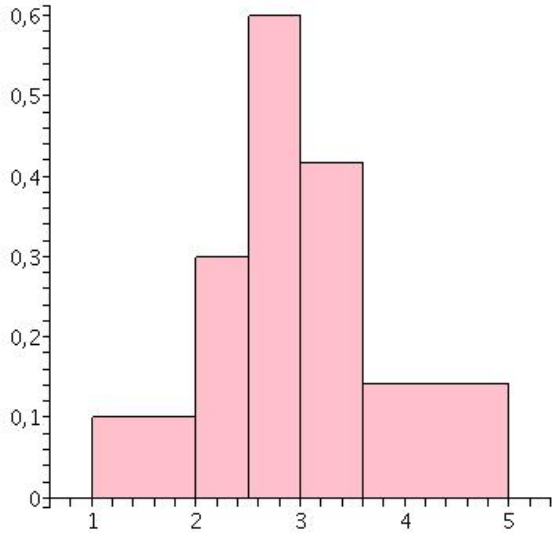
La técnica de Bootstrap fue introducida por B. Efron, 1982.

- Bootstrap: Levantarse tirando de las propias correas de las botas. Método autosuficiente.
- Fue diseñado para aproximar la precisión de un estimador a partir de un conjunto $x_1, x_2, \ldots, x_n$ de datos u observaciones.
- Determinar la precisión de muchos estimadores suele ser algebraicamente complicado, o imposible si no se conoce la distribución de los datos.
- El método Bootstrap permite obtener una buena aproximación del ECM (error cuadrático medio), del desvío y de la varianza de un estimador, a partir de la muestra, aún sin conocer la distribución de donde provienen los datos.

Técnica de bootstrap

La idea central de este método es simple;

- Dada una muestra aleatoria con *n* observaciones, dicha muestra es tratada como si fuera toda la población de las cuál se extraerán *B* muestras con reposición.
- Para cada una de las *B* nuevas muestras, se realizará una estimación del parámetro de interes.



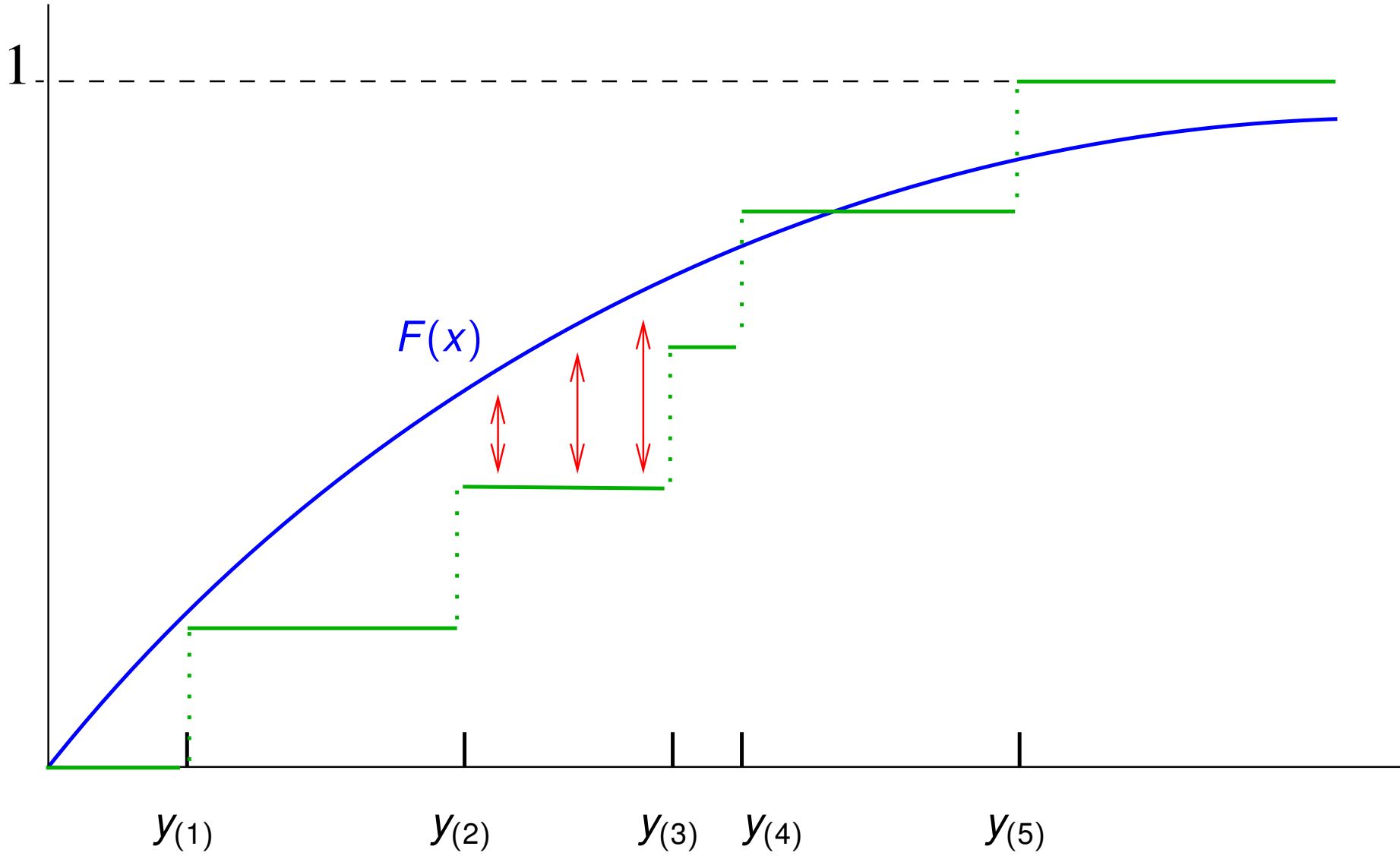
-

Se usarán los *B* valores bootstrap estimados para aproximar la distribución del estimador del parámetro, y funcionales derivados de dicha distribución (como la precisión).

Precaución

- Si la muestra original es una mala representación de la población en general, a pesar de ser una muestra aleatoria bien tomada, la simulación bootstrap va a generar una estimación pobre de la precisión.
- La estadística no mejora un diseño pobre.
- Los modelos deben ser elegidos con conocimiento general del problema, y luego ajustados con los datos recolectados para tal fin.

Distribución empírica de un grupo de datos fijo

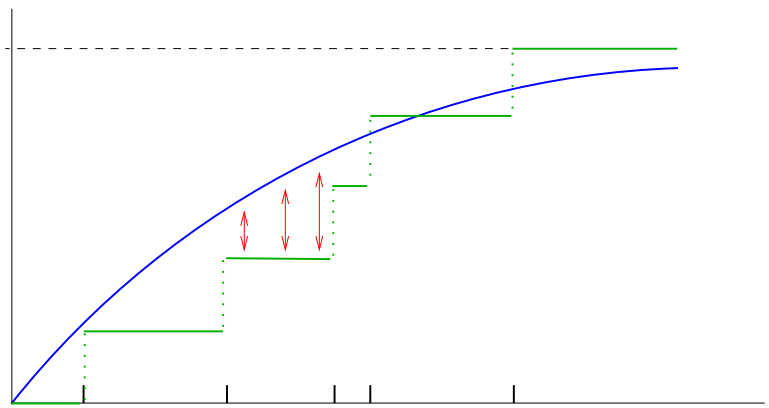


Distribución empírica de un grupo de datos fijo

Recordemos que si tenemos un grupo de datos observados: $x_1, x_2, \ldots, x_n$ disponibles, se define la distribución acumulada empírica F_e del grupo de datos

- Ordenando los datos de menor a mayor: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$
- Definiendo $F_e(x)$ como:

$$F_e(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{i}{n} & x_{(i)} \leq x < x_{(i+1)} \quad 1 \leq i < n \\ 1 & x_{(n)} \leq x \end{cases}$$

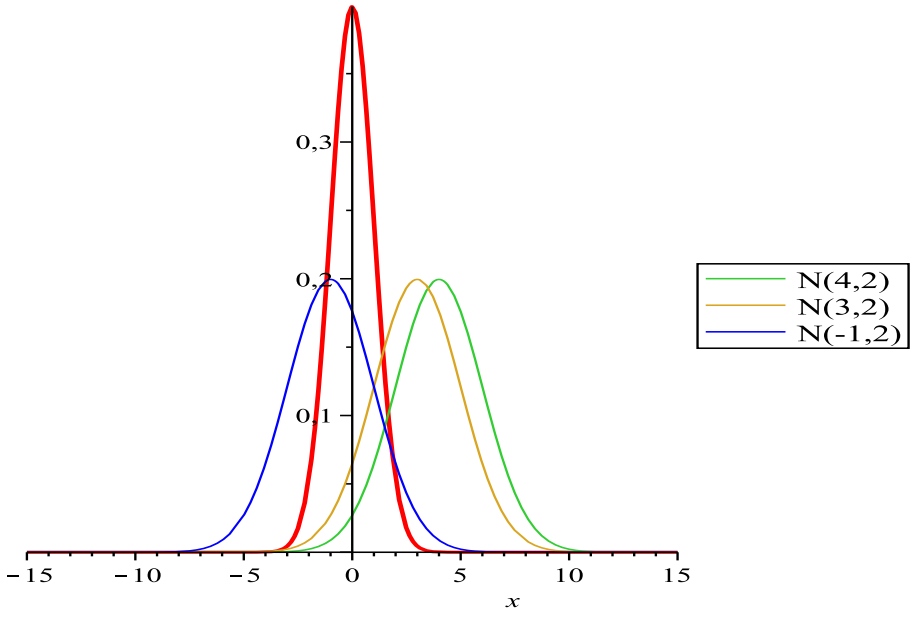


$F_e(x_{(i)}) = i/n \approx$ proporción de x_j menores que $x_{(i)}$.

Consideraciones estadísticas

Si el grupo de datos observados: $x_1, x_2, \ldots, x_n$ es una realización de una muestra aleatoria $X_1, X_2, \ldots, X_n$, esto es, v. a. independientes con la misma distribución F,

$$F(x) = P(X \leq x)$$



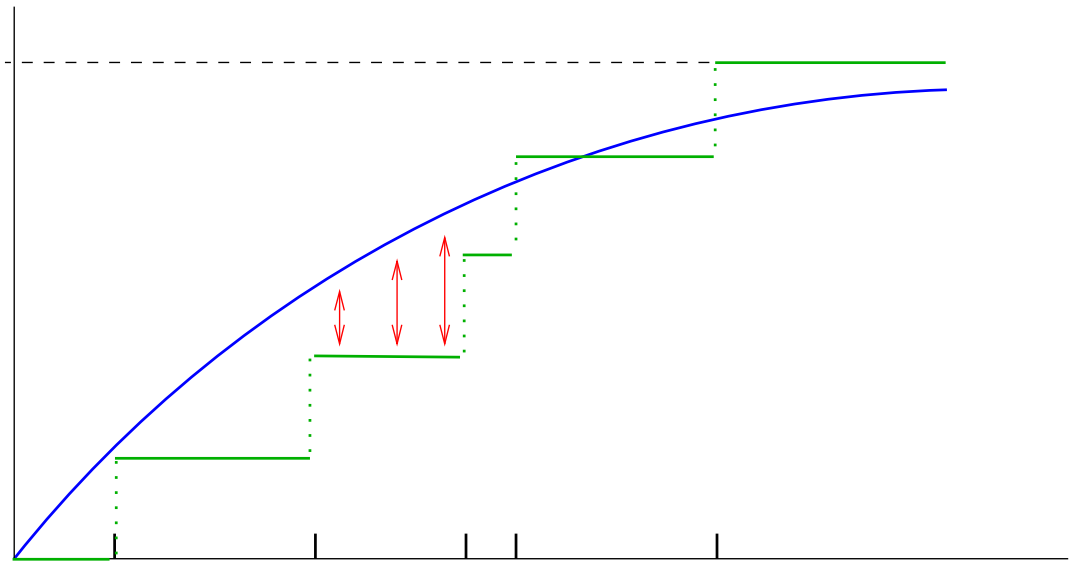
entonces los parámetros

$$\mu = E[X_i] \quad \sigma^2 = \mathrm{Var}(X_i)$$

dependen de la distribución F.

Distribución empírica

- Supongamos que el grupo de datos observados: $x_1, x_2, \ldots, x_n$ es una realización de una muestra aleatoria $X_1, X_2, \ldots, X_n$, esto es, v. a. independientes con la misma distribución F.
- La variable discreta X_e que toma los valores de la muestra $x_1, \ldots x_n$ con igual probabilidad $1/n$ (combinando los pesos si los x_i no son todos distintos) tiene distribución acumulada F_e,



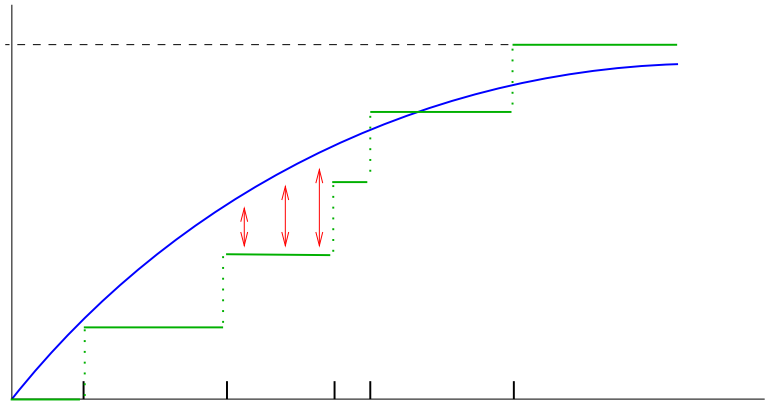
Distribución empírica

La variable discreta X_e tiene distribución F_e, por lo cual:

$$\mu_{F_e} = E_{F_e}[X_e] = \sum_{i=1}^{n} x_i P(X_e = x_i) = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X}$$

$$\sigma^2_{F_e} = \mathrm{Var}_{F_e}(X_e) = E_{F_e}[(X_e - \mu_{F_e})]^2 = \sum_{i=1}^{n} (x_i - \mu_{F_e})^2 P(X_e = x_i) = \sum_{i=1}^{n} \frac{(x_i - \mu_{F_e})^2}{n}$$

$$\sigma^2_{F_e} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{X})^2 = \frac{n-1}{n} S^2$$



También se definen otras medidas y parámetros: mediana, curtosis, coeficiente de asimetría, etc., todos ellos dependientes de la distribución.

Estimaciones

Si $X_1, X_2, \ldots, X_n$ son v.a. independientes, con distribución común F, tal que

$$E[X] = \mu, \quad \text{Var}(X) = E[(X-\mu)^2] = \sigma^2, \quad E[(X-\mu)^4] = \mu_4$$

La media muestral y la varianza muestral se definen como:

$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}(n))^2.$$

Estimaciones

$\overline{X}(n)$ y $S^2(n)$ son estimadores insesgados de μ y σ^2.

$$E[\overline{X}(n)] = \mu \qquad E[S^2(n)] = \sigma^2.$$

La varianza del estimador $\overline{X}(n)$ está dada por

$$\text{Var}(\overline{X}(n)) = \sigma^2/n.$$

La varianza del estimador $S^2(n)$ es

$$\text{Var}(S^2(n)) = \frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)}$$

Estimaciones

Distribución F

$$E[\overline{X}(n)] = \mu, \qquad E[S^2(n)] = \sigma^2, \qquad \text{Var}(\overline{X}(n)) = \sigma^2/n.$$

Distribución F_e

$$E_{F_e}[\overline{X}(n)] = \mu_{F_e}, \qquad E_{F_e}[S^2(n)] = \sigma^2_{F_e}, \qquad \text{Var}_{F_e}(\overline{X}(n)) = \sigma^2_{F_e}/n.$$

$$\mu_{F_e} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2_{F_e} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{F_e})^2$$

Técnica de bootstrap

- Así como $\overline{X}(n)$ y $S^2(n)$, pueden definirse otros estimadores para un determinado parámetro θ.
- Si $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$ es un estimador para un parámetro θ, interesa conocer

$$\text{Var}(\hat{\theta}) \qquad \text{y} \qquad ECM(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2].$$

- Estos valores suelen ser difíciles de calcular algebraicamente o numéricamente, más aún si no se conoce la distribución. Por ejemplo

$$ECM(\hat{\theta}, \theta) = \int \cdots \int (g(x_1, \ldots, x_n) - \theta)^2 \, f(x_1) \ldots f(x_n) dx_1 \ldots dx_n.$$

- La técnica de Bootstrap propone aproximar esta estimación utilizando la distribución empírica.

Técnica bootstrap

- Si n es suficientemente grande, suele ser cierto que:
  - (Glivenko-Cantelli): F_e converge uniformemente en x a F, con probabilidad 1.
  - Puede suponerse que los parámetros $\theta(F_e)$ de F_e se aproximan a los parámetros θ de F de manera continua.
- Entonces, por ejemplo: el error cuadrático medio del estimador $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$,

$$ECM(\hat{\theta}, \theta) = \int \cdots \int (g(x_1, \ldots, x_n) - \theta)^2 \, f(x_1) \ldots f(x_n) dx_1 \ldots dx_n.$$

podría aproximarse por:

$$ECM_e(\hat{\theta}, \theta) = E_{F_e}[(g(X_1, X_2, \ldots, X_n) - \theta(F_e))^2],$$

- $ECM_e(\hat{\theta}, \theta)$: aproximación bootstrap al error cuadrático medio.

Bootstrap

- Una aproximación bootstrap requiere una suma de n^n términos, si la muestra es de n observaciones.
- Cualquiera sea el estimador $\hat{\mu} = g(X_1, \ldots, X_n)$ de un parámetro μ, la estimación bootstrap de $ECM(\hat{\mu})$:

$$\sum_{1 \leq i_1 \leq n} \cdots \sum_{1 \leq i_n \leq n} \frac{(g(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) - \mu(F_e))^2}{n^n}$$

- Por ejemplo, para la aproximación de $ECM(S^2(n))$, se debe calcular:
  - μ_{F_e}: una vez.
  - Var_{F_e}: una vez.
  - Por cada una de las n^n muestras calcular el promedio $\overline{x}(n)$ y la varianza muestral $s^2(n)$ y hacer

$$(s^2(n) - \text{Var}_{F_e})^2.$$

Ejemplo

A partir de las 2 observaciones

$$X_1 = 1, \qquad X_2 = 3,$$

calcular la aproximación bootstrap de $ECM(\overline{X}, \mu)$ y $ECM(S^2, \sigma^2)$, siendo $\overline{X} = \frac{1}{2}(X_1 + X_2)$ y $S^2 = \frac{1}{2-1}\sum_{i=1}^{2}(X_i - \overline{X})^2$.

- Dado que $\overline{X}$ y S^2 son estimadores insesgados de la media y de la varianza respectivamente, se tiene que el error cuadrático medio con respecto a estos parámetros es igual a la varianza.
- $ECM(\overline{X}, \mu) = \text{Var}(\overline{X}) = E[(\overline{X} - E[\overline{X}])^2]$,
  $ECM(S^2, \sigma^2) = \text{Var}(S^2) = E[(S^2 - E[S^2])^2]$.
- Para la aproximación bootstrap utilizamos la distribución empírica, que da peso $p_1 = p_2 = \frac{1}{2}$.

Aproximación bootstrap

Varianza de la media muestral: $\mathrm{Var}(\overline{X})$

$$\mathrm{Var}_{F_e}(\overline{X}) = E_{F_e}[(\overline{X} - E_{F_e}[\overline{X}])^2] = \sum_{i=1}^{n^n} \frac{(\overline{X}_i - \mu_{F_e})^2}{n^n}$$

Observaciones originales:

$$X_1 = 1, \qquad X_2 = 3,$$

$E_{F_e}[\overline{X}] = \mu_{F_e} = \frac{1+3}{2} = 2.$

Muestras x_1	Muestras x_2	$\overline{X}$	$(\overline{X} - 2)^2$
1	1	1	1
1	3	2	0
3	1	2	0
3	3	3	1

$$\mathrm{Var}_{F_e}(\overline{X}) = \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 = \frac{1}{2}.$$

Aproximación bootstrap

Varianza de la varianza muestral: $\mathrm{Var}(S^2)$

$$\mathrm{Var}_{F_e}(S^2) = E_{F_e}[(S^2 - E_{F_e}[S^2])^2] = \sum_{i=1}^{n^n} \frac{(S_i^2 - \sigma_{F_e}^2)^2}{n^n}$$

Observaciones originales:

$$X_1 = 1, \qquad X_2 = 3,$$

$$E_{F_e}[S^2] = \sigma_{F_e}^2 = \frac{(1-2)^2+(3-2)^2}{2} = 1.$$

Muestras x_1	Muestras x_2	$\overline{X}$	S^2	$(S^2-1)^2$
1	1	1	0	1
1	3	2	2	1
3	1	2	2	1
3	3	3	0	1

$$\mathrm{Var}_{F_e}(S^2) = \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 1 = 1.$$

Bootstrap y Montecarlo

- Montecarlo: Este promedio puede aproximarse con un promedio de B términos, tomando B muestras aleatorias $(X_1^j, X_2^j, \ldots, X_n^j)$, $1 \leq j \leq B$:

$$\begin{aligned} Y_1 &= \left(g(X_1^1, X_2^1, \ldots, X_n^1) - \mu(F_e)\right)^2 \\ Y_2 &= \left(g(X_1^2, X_2^2, \ldots, X_n^2) - \mu(F_e)\right)^2 \\ &\vdots \\ Y_B &= \left(g(X_1^B, X_2^; \ldots, X_n^B) - \mu(F_e)\right)^2 \end{aligned}$$

$$ECM_e(\hat{\mu}) \approx \frac{\sum_{j=1}^{B} Y_j}{B}.$$

Aproximación bootstrap para $ECM(S^2, \sigma^2) = \text{Var}(S^2)$

- Obtener una muestra $x_1, x_2, \ldots, x_n$ (datos observados).
- F_e le asigna probabilidad $1/n$ a cada uno de estos datos (sumando pesos si no son todos distintos).
- Utilizar la distribución empírica para calcular:

$$E_{F_e}[(S^2(n) - E_{F_e}[S^2(n)])^2] = E_{F_e}[(S^2(n) - \text{Var}_{F_e}(X))^2]$$

siendo

$$\begin{aligned} \mu_{F_e} &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \\ \text{Var}_{F_e}(X) &= \frac{1}{n}\left((x_1 - \mu_{F_e})^2 + (x_2 - \mu_{F_e})^2 + \cdots + (x_n - \mu_{F_e})^2\right) \end{aligned}$$

Ejemplo

A partir de las 15 observaciones

$$5, 4, 9, 6, 21, 17, 11, 20, 7, 10, 21, 15, 13, 16, 8,$$

calcular la aproximación bootstrap de $\text{Var}(S^2) = \text{Var}(S^2(15))$.

- La distribución empírica da peso $p(21) = \frac{2}{15}$ y $p(x) = \frac{1}{15}$ a los restantes 13 valores.
- $\mu_{F_e} = 12.2$
- $\text{Var}_{F_e}(X) = 32.03$.
- Para cada una de las 15^{15} muestras $y_1, \ldots, y_{15}$ calcular
  - $\overline{y} = \frac{1}{15} \sum_{i=1}^{15} y_i$,
  - $s^2(n) = \frac{1}{14} \sum_{i=1}^{15} (y_i - \overline{y})^2$,
  - $(s^2(n) - 32.03)^2$,
- y promediar.

Consideraciones

Algorithm 1: X con distribución F_e

for $i = 1$ **to** nB **do**
  Generar $U \sim \mathcal{U}(0,1)$;
  $I \leftarrow \lfloor nU \rfloor + 1$;
  $X \leftarrow x[I]$
end

- Heurística: con $B = 100$ simulaciones se obtiene una buena aproximación de ECM_{F_e}.
- Esta aproximación bootstrap es a su vez una aproximación de ECM.

Aproximación bootstrap para $ECM(\overline{X}(n), \mu)$

El cálculo de $ECM(\overline{X}(n), \mu)$ es muy eficiente usando el valor observado de S^2/n, ya que

$$ECM(\overline{X}(n), \mu) = \text{Var}(\overline{X}(n)) = \sigma^2/n = E[S^2/n]$$

Pero veamos como se aproxima via bootstrap

- Si tenemos una muestra $x_1, x_2, \ldots, x_n$ de datos observados.
- Y F_e le asigna probabilidad $1/n$ a cada uno de estos datos (sumando pesos si no son todos distintos).
- El ECM estimado via Bootstrap resulta

$$ECM_{F_e}(\overline{X}(n), \mu_{F_e}) = E_{F_e}[(\overline{X}(n) - \mu_{F_e})^2].$$

Aproximación bootstrap para $ECM(\overline{X}(n), \mu)$

$$E_{F_e}(\overline{X}(n)) = \mu_{F_e}$$
$$\text{Var}_{F_e}\left(\overline{X}(n)\right) = \sigma_{F_e}/n$$

$$\begin{aligned} ECM(\overline{X}(n), \mu_{F_e}) &= E_{F_e}[(\overline{X}(n) - \mu_{F_e})^2] \\ &= E_{F_e}[(\overline{X}(n) - E_{F_e}(\overline{X}(n)))^2] \\ &= \text{Var}_{F_e}\left(\overline{X}(n)\right) \\ &= \frac{\sigma_{F_e}}{n} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \mu_{F_e})^2 \end{aligned}$$

Valor observado de S^2/n: $\dfrac{1}{n(n-1)}\displaystyle\sum_{i=1}^{n}(x_i - \mu_{F_e})^2$.

Ejemplo

Si se quiere estimar el tiempo promedio que un cliente pasa en un sistema debido a:

- Tiempo de espera en cola.
- Tiempo(s) de servicio.
- $W_i \leftarrow$ tiempo que permanece el i-ésimo cliente en el sistema.
- Se quiere calcular

$$\mu = \lim_{n \to \infty} \frac{W_1 + \cdots + W_n}{n}.$$

Ejemplo

- Notar: los tiempos W_i no son independientes ni idénticamente distribuidos.

En un caso simple de un solo servidor, en el que los clientes son atendidos por orden de llegada:

A_i: tiempo de arribo del cliente i.
S_i: tiempo de servicio del cliente i.
V_i: tiempo de salida del cliente i.

$$V_i = \max\{A_i, V_{i-1}\} + S_i, \qquad V_0 = 0$$

W_i: tiempo que pasa el cliente i en el sistema,

$$W_i = V_i - A_i = \max\{A_i, V_{i-1}\} + S_i - A_i.$$

$N_i \leftarrow$ número de clientes el día i:
$D_i \leftarrow$ suma de tiempos que permanecen los clientes en el sistema el día i:

$$\begin{aligned} D_1 &= W_1 + \cdots + W_{N_1} \\ D_2 &= W_{N_1+1} + \cdots + W_{N_1+N_2} \\ &\vdots \\ D_i &= W_{N_1+\cdots+N_{i-1}+1} + \cdots + W_{N_1+\cdots+N_i} \end{aligned}$$

- Notar: los tiempos D_i y los números N_i son independientes e idénticamente distribuidos.

Caracterización de μ

Por la ley de los grandes números

$$\begin{aligned}
\mu &= \lim_{n\to\infty} \frac{W_1 + \cdots + W_n}{n} \\
&= \lim_{m\to\infty} \frac{D_1 + \cdots + D_m}{N_1 + \cdots + N_m} \\
&= \lim_{m\to\infty} \frac{(D_1 + \cdots + D_m)/m}{(N_1 + \cdots + N_m)/m} \\
&= \frac{E[D]}{E[N]}
\end{aligned}$$

donde $E[N]$ es el número esperado de clientes que llegan en un día y $E[D]$ es el tiempo total esperado que esos clientes pasan en el sistema.

Estimación de μ

- Simular el sistema k días.
- Estimar $E[D]$ con $\overline{D} = \dfrac{D_1 + \cdots + D_k}{k}$.
- Estimar $E(N)$ con $\overline{N} = \dfrac{N_1 + \cdots + N_k}{k}$.
- Estimar μ con $\hat{\mu} = \dfrac{\overline{D}}{\overline{N}}$.

Para estimar

$$ECM\left(\frac{\overline{D}}{\overline{N}}\right) = E\left[\left(\frac{\sum_i D_i}{\sum_i N_i} - \mu\right)^2\right].$$

usamos el método bootstrap.

Aproximación bootstrap

- Observar valores d_i, n_i, $1 \leq i \leq k$.
- Distribución empírica:

$$P_{F_e}\left(D = d_i,\ N = n_i\right) = \frac{1}{k}$$

- $E_{F_e}(D) = \overline{d} = \sum_{i=1}^{k} d_i/k$.
- $E_{F_e}(N) = \overline{n} = \sum_{i=1}^{k} n_i/k$.
- $\mu_{F_e} = \dfrac{\overline{d}}{\overline{n}}$.

$$ECM_{F_e}\left(\frac{\overline{D}}{\overline{N}}\right) = \frac{1}{k^k} \sum_{(i_1,\ldots,i_k)} \left(\frac{d_{i_1} + \cdots + d_{i_k}}{n_{i_1} + \cdots + n_{i_k}} - \frac{\overline{d}}{\overline{n}}\right)^2.$$

Bootstrap aproximado

Este promedio puede aproximarse con un promedio de B términos, tomando B muestras aleatorias (D_i^j, N_i^j),$1 \leq i \leq k$, $1 \leq j \leq B$:

$$
\begin{aligned}
Y_1 &= \left(\frac{\sum_{i=1}^{k} D_i^1}{\sum_{i=1}^{k} N_i^1} - \frac{\overline{d}}{\overline{n}} \right)^2 \\
Y_2 &= \left(\frac{\sum_{i=1}^{k} D_i^2}{\sum_{i=1}^{k} N_i^2} - \frac{\overline{d}}{\overline{n}} \right)^2 \\
&\vdots \\
Y_B &= \left(\frac{\sum_{i=1}^{k} D_i^B}{\sum_{i=1}^{k} N_i^B} - \frac{\overline{d}}{\overline{n}} \right)^2
\end{aligned}
$$

$$
ECM_e(\hat{\mu}) \approx \frac{\sum_{j=1}^{B} Y_j}{B}.
$$

Ejercicio 8. Considerar un sistema con un único servidor en el cual los clientes potenciales llegan de acuerdo con un proceso de Poisson de razón 4.0.
Un cliente potencial entrará al sistema sólo si hay tres o menos clientes en el sistema al momento de su llegada.
El tiempo de servicio de cada cliente está distribuído según una exponencial de parámetro 4.2.
Despues del instante $T = 8$ no entran mas clientes al sistema (los tiempos están dados en horas).
Realizar un estudio de simulación para estimar el tiempo promedio que un cliente pasa en el sistema.
Aplicar el método “bootstrap” para estudiar el error cuadrático medio de su estimador.