Pro Mathematica: Vol. IX, Nos. 17-18, 1995

LOS PROCESOS EMPIRICOS Y EL METODO BOOTSTRAP

Luis ***Valdivieso***

Introducción

Los procesos empíricos son aquellos asociados a la llamada función de distribución empírica F_n *. Esta última debe fundamentalmente su importancia a su convergencia a la verdadera función de distribución F, razón por la cual muchos estadísticos de la forma* $\hat{\theta} = g(F_n)$ *son utilizados como estimadores de parámeteros* $\theta = g(F)$*.*

En relación a esta estimación ha surgido últimamente con gran impulso el estudio del denominado método bootstrap.

El presente artículo es una panorámica mirada a los procesos empíricos y a una de sus múltiples aplicaciones en la estimación por este método.

✏ *Profesor del Departamento de Ciencias, Sección Matemáticas, PUCP.*

Los Procesos Empíricos

Sean $X_1, X_2, ..., X_n$ variables aleatorias independientes e idénticamente distribuídas (i.i.d) sobre un espacio probabilístico $(\Omega, \mathcal{F}, P)$ y sea F la función de distribución de estas variables. La función de distribución empírica (f.d.e) está definida como:

$$(1) \qquad F_n(x) \equiv F_n(x,\omega) = \frac{\sum_{i=1}^{n} 1|_{\{X_i \le x\}}(\omega)}{n} = \frac{\sum_{i=1}^{n} 1|_{\{X_{(i)} \le x\}}(\omega)}{n}$$

donde

$$1|_{\{X_i \le x\}}(\omega) = \begin{cases} 1, & \text{si} \quad X_i(\omega) \le x \\ 0, & \text{en} \quad \text{otro caso} \end{cases}$$

Si fijamos un punto $x \in \mathbf{R}$, $nF_n(x) \sim$ Binomial $(n, F(x))$, lo cual implica por (1) y la ley fuerte de los grandes números (LFGN) que:

$$(2) \qquad F_n(x) \xrightarrow{c.p} F(x) \qquad \text{(c.p = casi en todas partes)}$$

Más aún por el Teorema de Límite Central (TLC), se tiene que:

$$(3) \qquad \sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} Y \sim N(0, F(x)(1-F(x))).^{1}$$

Estos análisis preliminares fueron hechos para un x particular. Aquí es interesante notar que $F_n(.)$ es una función aleatoria (depende de $\omega \in \Omega$). En consecuencia, podemos considerar a $\{F_n(x)\}_{x \in \mathbf{R}}$ como un proceso estocástico, cuyas realizaciones $\{F_n(.,\omega)\}$ (ω fijo) son continuas por la derecha y con límites por la izquierda. En general un proceso cuyas realizaciones comparten estas propiedades es llamado *cadlag* y al espacio de las funciones cadlag, ha de notarse por D, se le llama el espacio de Skorohod.

[1] *Si (X_n) y X son variables aleatorias que asumen valores en un espacio métrico M,*

$$X_n \xrightarrow{d} X \Leftrightarrow \forall f \in C(M), \quad E(f(X_n)) \to E(f(X)),$$

donde

$$C(M) = \{f : M \to R \mid f \text{ es medible continua y acotada}\}.$$

Esta definición es más general que la usual de convergencia de funciones de distribución $F_{X_n}(x) \to F_X(x)$ para puntos x de continuidad de F_X, y es equivalente a la anterior, sólo en el caso que M tenga dimensión finita.

El espacio D puede estar provisto de la métrica uniforme $||f||_\infty = \sup\{|f(t)| / t \in \mathbf{R}\}$ o de la métrica de Skorohod ([1]). Asimismo, para que D sea un espacio medible, pueden considerarse las σ-álgebras topológicas (de Borel), la generada por bolas cerradas o la proyectiva. Nosotros, salvo indicación contraria, trabajaremos con la métrica uniforme y la σ-álgebra de Borel.

Es de interés, dada la convergencia puntual de F_n a F, preguntarse si esta convergencia es uniforme. La respuesta es afirmativa y el resultado conocido como el teorema de Glivenko-Cantelli:

$$(4) \qquad \|F_n - F\|_\infty \xrightarrow{c.p} 0.$$

Este resultado puede interpretarse como la LFGN aplicado a las i.i.d. funciones cadlag, $1|_{\{X_i \le .\}}$ $(i = 1,2,\ldots,n)$. En efecto, F_n es por (1) justamente la medida aritmética de las funciones (elementos aleatorios) cadlag:

$$1|_{\{X_1 \le .\}}, 1|_{\{X_2 \le .\}}, \ldots, 1|_{\{X_n \le .\}}.$$

Esta analogía entre la convergencia del promedio de variables aleatorias i.i.d en (2) y la de elementos aleatorios i.i.d en (4), nos lleva a la consideración de una posible analogía para (3); es decir de un "TLC uniforme". El candidato natural para esta generalización del TLC es, según (3), $E_n = \sqrt{n}(F_n - F)$, tal proceso en el espacio D es llamado el proceso empírico de la muestra aleatoria $X_1, X_2, \ldots, X_n$.[2]

Teorema (Donsker) $E_n \xrightarrow{d} E = BoF$, *donde B es un puente Browniano*[3] *en* $[0,1]$ *con*

$$E(B(t)) = 0 \text{ y } \operatorname{cov}(B(t), B(s)) = s(1-t), \quad 0 \le s \le t \le 1.$$

Este teorema es llamado también el teorema del límite central funcional del proceso empírico.

[2] *El proceso empírico no es medible con la σ-álgebra de Borel. Por esta razón en el teorema de Donsker, a D se le dotará de la σ-álgebra proyectiva, en la cual* E_n *si es medible.*

[3] *Si* $\{W(t)\}_{t\in[0,1]}$ *es un movimiento Browniano en* [0,1], *al proceso B definido por* $\{B(t) = W(t) - tW(1)\}_{t\in[0,1]}$ *se le llama un puente Browniano en* [0,1].

El término funcional se acuña en razón de que por el teorema de la función continua en la teoría de convergencia estocástica ([6]), si f es una función continua de $D[-\infty,\infty]$ en $\mathbf{R}$, entonces $f(E_n) \xrightarrow{d} f(BoF)$.[4] Asi, por ejemplo:

$$f_1(E_n) = ||E_n||_\infty \xrightarrow{d} f_1(E) = ||BoF||_\infty$$

$$f_2(E_n) = \int E_n^2(t)dt \xrightarrow{d} f_2(E) = \int E^2(t)dt$$

Nótese que la primera expresión da respuesta a nuestra interrogante sobre la analogía "uniforme" para (3).

El TLC admite básicamente versiones menos generales que la arriba indicada, y más generales que el TLC clásico aplicado en (3). Entre estos destacan los TLC de Lyapunov, Lindeberg y el TLC multivariado. El TLC de Lyapunov, por citar un ejemplo -ya que lo utilizaremos en el estudio del método bootstrap- dice lo siguiente (5): Sea $(X_{ni})_{i=1,2,...k_n}$ una secuencia de variables aleatorias independientes y sea $S_n = X_{n1} + X_{n2} + ... + X_{nk_n}$, donde podemos asumir, sin pérdida de generalidad, que las variables aleatorias X_{ni} tienen media cero y varianza σ_{ni}^2 tales que

$$\sigma_{n1}^2 + \sigma_{n2}^2 + ... + \sigma_{nk_n}^2 = 1.$$

Entonces, si $\sum_{i=1}^{k_n} E|X_{ni}|^3 \to 0$ (condición de Lyapunov),

$$S_n \xrightarrow{d} Z \sim N(0,1).$$

Está por demás decir la cantidad de estudio que involucra un análisis más detallado de los proceso empíricos, en especial por ejemplo entender el significado de la convergencia en el teorema de Donsker. Nosotros haremos un alto aqui y mostraremos para mejor comprensión una de las aplicaciones de esta clase de estudios en el llamado método bootstrap.

[4] *En general, el teorema afirma que, si M_1 y M_2 son dos espacios métricos y medibles, $f: M_1 \to M_2$ una función medible y C el conjunto de puntos de continuidad de f, entonces $X_n \xrightarrow{d} X$ en M_1, donde $P(X \in C) = 1 \Rightarrow f(X_n) \xrightarrow{d} f(X)$ en M_2.*

El Método Bootstrap

Sea $\underline{X} = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria con función de distribución F y sea $\theta = g(F)$ un parámetro a estimarse. Un posible estimador de θ podría ser, como se expresó en la introducción, $\hat{\theta} = g(F_n) = h(\underline{X})$. La cuestión radica en cuán eficiente es este estimador. El método bootstrap fue introducido por Efron en 1979 como un método computacional para la estimación del error estándar de $\hat{\theta}$. Esta estimación no requiere de cálculos teóricos, es independiente de cuán complicada sea la función h y es no paramétrica, vale decir, independiente de la asunción de un modelo para F. Todo esto es factible pues sabemos por lo desarrollado que F puede aproximarse por la f.d.e F_n.

El método bootstrap depende de la noción de muestra bootstrap: Si F_n es la f.d.e que asigna probabilidad 1/n a cada valor observado de la muestra aleatoria (ver (1) para mayor formalidad). Una muestra bootstrap $\underline{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ es una selección de n elementos con reemplazamiento de $\underline{X}$, o desde otra perspectiva, una muestra aleatoria de tamaño n de la "población" $\underline{X}$, con f.d.e F_n. A $\hat{\theta}^* = h(\underline{X}^*)$ se le llama la réplica bootstrap de $\hat{\theta}$, y desde que nuestro interés es estimar $se_F(\hat{\theta}) = \sqrt{\text{var}_F(\hat{\theta})}$ para medir la eficiencia de $\hat{\theta}$, se define su estimador ideal bootstrap, mediante:

(6) $$se_{F_n}(\hat{\theta}^*).$$

Con respecto a esta definición, cabe hacer dos observaciones importantes:

a) Si por ejemplo $\theta = \mu = g(F) = E_F(X)$ y $\hat{\theta} = \bar{X} = \sum_{i=1}^{n} \frac{1}{n} X_i = g(F_n)$, entonces $se_F(\hat{\theta}) = \sqrt{\text{var}_F(X)/n}$. Esta formulación nos lleva a pensar, porque en vez de utilizar (6) no estimamos tan solo F por F_n y consecuentemente obtenemos el estimador $se_{F_n}(\hat{\theta}) = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\ / n$. El problema de este razonamiento estriba en que si $\hat{\theta}$ no es la media muestral, difícilmente podrá encontrarse una formulación exacta como la anterior para

$se_F(\hat{\theta})$; imagínese por ejemplo los casos en que $\hat{\theta}$ = Mediana muestral o $\hat{\theta}$ = Coeficiente de correlación muestral, como estimadores de sus correspondientes parámetros poblacionales.

b) Si bien es posible calcular teóricamente $se_{F_n}(\hat{\theta}^*)$, el procedimiento es completamente impráctico en términos computacionales, debido básicamente a la gran cantidad $m = \binom{2n-1}{n} = \frac{(2n-1)!}{(n-1)!\,n!}$ de distintas muestras bootstrap existentes.

Es en razón de los problemas en a) y b) que se ha implementado el siguiente procedimiento para obtener una buena aproximación de $se_{F_n}(\hat{\theta}^*)$:

(7.1) Seleccionar B muestras independientes bootstrap $\underline{X}^{*1}, \underline{X}^{*2}, \ldots, \underline{X}^{*B}$, cada una consistente de n extracciones con sustitución de $\underline{X}$.

(7.2) Evaluar la réplica bootstrap de $\hat{\theta}$ para cada muestra bootstrap:

$$\hat{\theta}^*(b) = h(\underline{X}^{*b}), \quad b = 1,2,3,\ldots,B.$$

(7.3) Estimar $se_{F_n}(\hat{\theta}^*)$ por la desviación estandar muestral de las B réplicas:

$$s\hat{e}_B = \sum_{b=1}^{B} (\hat{\theta}^*(b) - \hat{\theta}^*(.))^2 / (B-1), \text{ donde } \hat{\theta}^*(.) = (\sum_{b=1}^{B} \hat{\theta}^*(b)) / B.$$

$s\hat{e}_B$ es llamado el estimador bootstrap del error estandar de $\hat{\theta}$.

Dado que $\lim_{B\to\infty} s\hat{e}_B = se_{F_n}(\hat{\theta}^*)$, en la práctica suele tomarse B entre 25 y 200, aunque a veces mucho más grande, dependiendo concretamente del tipo de estimación.

Una excelente introducción al método bootstrap puede encontrarse en ([2]), nosotros nos limitaremos ahora al caso $\theta = \mu$ y $\hat{\theta} = \overline{X}$, y probaremos teóricamente que el método bootstrap "funciona" para el caso de la media.

Para un estudio de otros estadísticos uno puede remitirse a Gill ([3]) o a Hall ([4]). Ambos autores, utilizan como herramienta básica el TLC de Donsker reseñado.

El TLC Bootstrap

Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d con función de distribución F, media μ y varianza σ^2, y sea F_n la f.d.e de los primeros n elementos. Dado F_n consideremos la muestra bootstrap $\underline{X}^* = (X^*_{n1}, X^*_{n2}, \ldots, X^*_{nn})$ y sea $P^*(.) = P(. / X_1, X_2, \ldots, X_n)$. Denotaremos con E^* y Var^* a la media y varianza con respecto a P^*. Nosotros deseamos aplicar el TLC a la suma bootstrap:

$$S^*_n = X^*_{n1} + X^*_{n2} + \ldots + X^*_{nn}.$$

Dado que

$$E^*(X^*_{ni}) = \sum_{i=1}^{n} \frac{1}{n} X_i = \overline{X} \quad y \quad Var^*(X^*_{ni}) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n} = SS^2,$$

podríamos considerar el arreglo doble (condicional)

$$\left(Z^*_{ni} = \frac{X^*_{ni} - E^*(X^*_{ni})}{\sqrt{n\, Var^*(X^*_{ni})}} = \frac{X^*_{ni} - \overline{X}}{\sqrt{n\, SS^2}} \right)_{i=1,2,\ldots,n}$$

que presenta todas las características de (5). Resta verificar la condición de Lyapunov. En efecto, desde que es posible probar en base al teorema de Borel-Cantelli y la LFGN que $\max_i \frac{|X_i - \overline{X}|^2}{n} \xrightarrow{c.p} 0$, $SS^2 \xrightarrow{c.p} \sigma^2$ y,

$$\sum_{i=1}^{n} E^*|Z_{ni}^*|^3 = \frac{1}{\sqrt{n}\,SS^3} E^*|X_{ni}^* - \overline{X}|^3 = \frac{1}{\sqrt{n}\,SS^3}\frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}|^3 \leq$$

$$\leq \frac{1}{\sqrt{n}\,SS} \max_{i=1,2,\ldots,n} |X_i - \overline{X}|$$

la condición de Lyapunov es satisfecha. Así, aplicando el TLC respectivo y el teorema de la función continua, se obtiene que:

$$\sup_{x\in\mathbf{R}} \left| P^*\left(\frac{S_n^* - n\,E^*(X_{n1}^*)}{\sqrt{Var^*(S_n^*)}} \leq x \right) - \phi(x) \right| \longrightarrow 0 \ldots (*).^5$$

Por otra parte, aplicando el clásico TLC y el teorema de la función continua, se obtiene también que:

$$\sup_{x\in\mathbf{R}} \left| P\left(\frac{S_n - n\,\mu}{\sqrt{n\sigma^2}} \leq x \right) - \phi(x) \right| \longrightarrow 0 \ldots (**)$$

Así, una combinación de (*) y (**) lleva a afirmar que el método bootstrap funciona para el caso de la media, esto es:

$$\sup_{x\in\mathbf{R}} \left| P^*\left(\frac{\overline{X}_n^* - E^*(X_{n1}^*)}{\sqrt{\frac{Var^*(X_{n1}^*)}{n}}} \leq x \right) - P\left[\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq x \right] \right| \longrightarrow 0$$

para casi toda realización de (X_n).

[5] ϕ *es la función de distribución de* $Z \sim N(0,1)$.

Bibliografía

[1] *Billingsley, Patrick:* Convergence of Probability Measures, Wiley. New York. 1968.

[2] *Efron, Bradley and Tibshirani:* An Introduction to the Bootstrap. Chapman and Hall. 1993.

[3] *Gill,Richard:* Non- and Semi-parametric Maximum Likelihood Estimators and the Von Mises Method (Part 1). Scandinavian Journal Statistics. 1989.

[4] *Hall, P.:* The Bootstrap and Edgeworth Expansions. Springer. New York. 1992.

[5] *Mikosh, Thomas:* Empirical Processes's lectures notes M.R.I. The Netherlands. 1994.

[6] *Pollard, David:* Convergence of Stochastic Processes. Springer-Verlag. 1984.

lvaldiv@pucp.edu.pe