

**ESTIMACION NO PARAMETRICA DE LA FUNCION DE
DENSIDAD DE TRANSICION DE UN PROCESO DE MARKOV
(VIA BOOTSTRAP)**

JORGE ALI MENDEZ FERNANDEZ

TRABAJO DE GRADO

Presentado ante la

Universidad de los Andes

En cumplimiento parcial de los requisitos

para optar al título de

MAGISTER SCIENTIARUM

Instituto de Estadística Aplicada y Computación

Mérida, Diciembre 1994

AGRADECIMIENTO

Quiero expresar mi agradecimiento a los profesores Elizabeth Torres, Arnaldo Goitia y en especial la estrecha colaboración y asesoría del Prof. Jose Manuel Hernández C., Profesor Invitado, Universidad de La Habana - Cuba, en la realización de este trabajo.

INDICE

CAPITULO I	
1.- INTRODUCCION	1
CAPITULO II	
2.- ANTECEDENTES	5
CAPITULO III	
3.- METODOLOGIA ESTADISTICA	14
3.1.- Método Bootstrap	
3.1.1.- Introducción	
3.1.2.- Método Bootstrap	18
3.1.2.1.- Algoritmo del Método Bootstrap	19
3.1.3.- Errores Estándares y Estimadores de Errores Estándares	20
3.1.3.1.- Introducción	
3.1.3.2.- El Estimador Bootstrap del Error Estándar	
3.1.3.3.- El Número de Replicaciones Bootstrap	22

3.1.3.4.- Estructuras de Datos más complicadas	23
3.1.4.- Estimadores del Sesgo	26
3.1.5.- Propiedades Asintóticas del Método Bootstrap	27
3.1.6.- Intervalos Confidenciales Bootstrap	28
3.1.6.1.- El Intervalo <i>t-bootstrap</i>	29
3.1.6.2.- El Intervalo Percentil	30
3.1.6.3.- El Método <i>Bias-Corrected and Accelerated (BCa)</i>	31
 3.2.- Método Kernel	34
3.2.1.- Introducción	
3.2.2.- Estimadores de Densidad Confiables	36
3.2.3.- Propiedades Estadísticas de los Estimadores de Densidad	37
3.2.4.- Estimadores Kernels	40
3.2.4.1.- Conceptos Informales de Estimación Kernel	
3.2.4.2.- Propiedades Estadísticas Fundamentales	45
3.2.4.3.- Las Medidas <i>MSE</i> y <i>MISE</i>	50
3.2.5.- Selección del Parámetro de Suavización	51
3.2.5.1.- Introducción	
3.2.5.2.- Lista de Métodos	52
3.2.6.- Selección de un Kernel	56
3.2.7.- Intervalos Confidenciales para la Estimación No Paramétrica de Densidades	58
3.2.7.1.- Introducción	
3.2.7.2.- Diferentes Métodos Bootstrap en la Estimación No Paramétrica de densidades	59

3.2.7.3.- Intervalos Confidenciales para $\hat{E}\hat{f}(x), \hat{f}(x)$	65
3.3.- Análisis de Resultados de las Simulaciones	67
3.3.1.- Procesos de Markov y sus Distribuciones Teóricas	
3.3.2.- Análisis de los Resultados	70
3.3.3.- Conclusiones	83
BIBLIOGRAFIA	92
ANEXO A: Implementación Computacional	95
ANEXO B: Gráficos Adicionales	114

QA 274.7
M216

SUMARIO

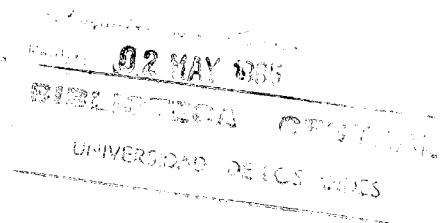
ESTIMACION NO PARAMETRICA DE LA FUNCION DE DENSIDAD DE TRANSICION DE UN PROCESO DE MARKOV (VIA BOOTSTRAP)

por

JORGE ALI MENDEZ FERNANDEZ

El método kernel fue utilizado para estimar la función de densidad de transición de Procesos de Markov (estacionarios), en el caso gaussiano y algunos casos no gaussianos en particular. Además el método bootstrap fue utilizado para estudiar el comportamiento de las estimaciones obtenidas a partir del método kernel mediante bandas confidenciales (puntuales) asintóticas y bootstrap, y como un posible mejoramiento de las mismas.

La implementación computacional se hizo en SAS/IML versión 6.09.



CAPITULO I

1.- INTRODUCCION

(Skorohod, 1977) La idea de un Proceso "sin efecto posterior" es la característica fundamental de un Proceso de Markov. Considérese un sistema (o una partícula) que puede encontrarse en varios estados. Los posibles estados forman un conjunto A que se denomina espacio fase del sistema. Asumiendo que el sistema cambia en el tiempo, el estado del sistema en el tiempo t es denotado por x_t . Si $x_t \in B$ y $B \subseteq A$, decimos que el sistema en el tiempo t está situado en el conjunto B .

Asumiendo que la evolución del sistema es de naturaleza estocástica, es decir, el estado del sistema en el tiempo t , en general, no está únicamente determinado por el estado del sistema en el tiempo s , donde $s < t$, sino que es aleatorio y está descrito por ciertas leyes probabilísticas. Denotaremos por $P(s, x, t, B)$ la probabilidad del evento $x_t \in B$ ($s < t$) / $x_s = x$. La función $P(s, x, t, B)$ se denomina la probabilidad de transición del sistema dado. Un sistema es denominado "sin efecto posterior" si la probabilidad de

que estando situado en el tiempo t en el conjunto B , bajo la condición de que el movimiento del sistema en el tiempo s ($s < t$) es completamente conocida, es igual a $P(s, x, t, B)$ y éste sólo depende del estado del sistema en el tiempo s .

Denotemos por $P(s, x, u, y, t, B)$ la probabilidad condicional del evento $x_t \in B$ bajo los supuestos $x_s = x, x_u = y, (s \leq u \leq t)$. De las propiedades de la esperanza condicional se tiene que

$$P(s, x, t, B) = \int_A P(s, x, u, y, t, B) P(s, x, u, dy) \quad (1.1)$$

Para un sistema sin efecto posterior $P(s, x, u, y, t, B) = P(u, y, t, B)$, en este caso (1.1) nos queda,

$$P(s, x, t, B) = \int_A P(u, y, t, B) P(s, x, u, dy) \quad (1.2)$$

La ecuación (1.2) se denomina *ECUACION DE CHAPMAN-KOLMOGOROV*. Esto puede servir como base para una definición de un proceso sin un efecto posterior, tales procesos se conocen con el nombre de *PROCESOS DE MARKOV*. //

(Doob, 1967) Un Proceso de Markov es un proceso $\{X_t : t \in T\}$ que satisface la condición siguiente: para cualquier entero $n \geq 1$, si t_1, \dots, t_n son valores paramétricos, la probabilidad condicional de X_{t_n} conocidos los valores de $X_{t_1}, \dots, X_{t_{n-1}}$ es la misma probabilidad condicional de X_{t_n} conocido sólo el valor de $X_{t_{n-1}}$, en el sentido de que para cada α ,

$P[X_{t_n}(w) \leq \alpha / X_{t_1}, \dots, X_{t_{n-1}}] = P[X_{t_n}(w) \leq \alpha / X_{t_{n-1}}] \quad (1.3)$
con probabilidad uno. //

(Karlin et al, 1974) Un Proceso de Markov, es un proceso que estrictamente hablando, tiene la propiedad de que dado el valor de X_t , los valores $X_s, s > t$, no dependen de los valores de $X_u, u < t$, es decir, las probabilidades de cualquier conducta futura del proceso, cuando su estado

presente se conoce exactamente, no es alterado por el conocimiento adicional de la conducta del proceso en el pasado.

Dada una realización en particular del proceso $\{X_t : t = 1, \dots, N\}$, podemos realizar inferencias acerca del mismo, nuestro interés particular se encuentra en la estimación no paramétrica de la función de probabilidad de transición P , la cual, en general, es desconocida, ¿Cómo vamos a estimar a P ?

La estimación no paramétrica de densidades es un interesante tópico la cual se ha desarrollado en años recientes, debido en parte a los modernos computadores y al desarrollo de gráficos de alta resolución, constituye una herramienta alterna para el análisis de datos univariantes y multivariantes en grandes proporciones.

Nuestra hipótesis es que mediante el método bootstrap se puede estudiar y mejorar el comportamiento de uno de los métodos no paramétricos de estimación de densidades, en particular, el *Método Kernel*, para estimar la función de densidad de transición de un Proceso de Markov, considerando la estructura estadística del proceso: gaussiana o no gaussiana, tomando en cuenta algunos estimadores ya desarrollados en la literatura, a través de bandas confidenciales (puntuales) bootstrap.

Los objetivos principales de este trabajo son:

Generales:

Estudiar el comportamiento (vía bootstrap) de la estimación de la f.d.p. de transición, a través del método kernel, en el caso normal y algunos casos no normales en particular: exponencial, beta, uniforme continua.

Específicos:

- (i) Revisión Bibliográfica
- (ii) Implementación computacional
- (iii) Estudiar el comportamiento del Método Kernel para diferentes tamaños de muestras.
- (iv) Estimar la varianza del estimador de la densidad a través del método bootstrap.
- (v) Establecer intervalos de confianza bootstrap (puntuales), como una alternativa a los intervalos confidenciales teóricos.
- (vi) Realizar comparaciones, entre los intervalos confidenciales estándares con la varianza estimada y los intervalos bootstrap, para distintos casos.

CAPITULO II

2.- ANTECEDENTES

(*Masry, 1989*) Sea $\{X_j : j \in \mathfrak{N}\}$ un proceso estacionario real valorado sobre el espacio de probabilidad (Ω, \mathcal{A}, P) . Dada una realización en particular $\{X_j : j = 1, \dots, n\}$ del proceso, pueden realizarse inferencias acerca del mismo. De interés particular es la estimación no paramétrica de la f.d.p. de transición.

Para n "grande" tales estimadores para procesos estacionarios combinados pueden obtenerse claramente a partir de la estructura estadística del proceso, esto es, considerando su naturaleza gaussiana o no gaussiana.

Para cada entero $m \geq 1$ y enteros $0 = i_1 < i_2 < \dots < i_m$. Sea $f(x, im) = f(x_1, \dots, x_m; i_1, \dots, i_m)$ la función de densidad de probabilidad conjunta de las variables aleatorias X_{i_1}, \dots, X_{i_m} , la cual se asume que existe. La función de densidad de probabilidad condicional de

$X_j'' = (X_{j+i_p+1}, \dots, X_{j+i_m})$ dado $X'_j = (X_{j+1}, \dots, X_{j+i_p})$ esta dada por: $f(x_2/x_1) \cong \frac{f(x, i_m)}{f(x_1, i_p)}$,

donde $x_1 \in \mathbb{R}^p, x_2 \in \mathbb{R}^{m-p}$ y $x = (x_1, x_2) \in \mathbb{R}^m$. Sea q una función Borel medible sobre \mathbb{R}^{m-p} tal que $E[|q(X_j'')|] < \infty$.

Para cada $l \geq 1$, sea $\mathbf{K}_l(x)$ una función acotada no negativa sobre \mathbb{R}^l que satisface las condiciones siguientes:

$$(i) \int_{\mathbb{R}^l} \mathbf{K}_l(u) du = 1; \quad \mathbf{K}_l(x) = O(\|x\|)^{-1-\xi}, \quad \xi > 0.$$

Sea $\{b_j : j \geq 1\}$ una sucesión de números positivos tal que $b_j \rightarrow 0$ cuando $j \rightarrow \infty$, y hacemos $\mathbf{K}_{l,j}(x) = \left(\frac{1}{b_j^l}\right) \mathbf{K}_l\left(\frac{x}{b_j}\right)$.

En base a una realización particular $\{X_j : j = 1, \dots, n\}$ se estima $\hat{\mathbf{f}}(x, i_m), m \geq 1$, por

$$\hat{\mathbf{f}}_n(x, i_m) = (n - i_m)^{-1} \sum_{j=1}^{n-i_m} \mathbf{K}_{m,j}(x - X_j) \quad (2.1)$$

donde, $X_j = (X_{j+i_1}, \dots, X_{j+i_m})$ y naturalmente se asume que $n > i_m$.

El estimador (2.1) es claramente recursivo ya que

$$\hat{\mathbf{f}}_n(x, i_m) = \left(\frac{n-1-i_m}{n-i_m}\right) \hat{\mathbf{f}}_{n-1}(x, i_m) + (n - i_m)^{-1} \mathbf{K}_{n-i_m}(x - X_{n-i_m}).$$

La f.d.p. condicional $f(x_2/x_1)$ se estima mediante la expresión siguiente,

$$\hat{\mathbf{f}}_n(x_2/x_1) = \frac{\hat{\mathbf{f}}_n(x_2, i_m)}{\hat{\mathbf{f}}_n(x_1, i_p)} \quad (2.2)$$

y cada uno de los estimadores antes mencionados tienen convergencia casi segura a su verdadero valor. //

(*Roussas, 1969*) En un proceso de Markov fundamental, éste define un espacio de probabilidad (Ω, \mathcal{A}, P) y toma valores en la recta real, es

(estRICTAMENTE) estacionario y tiene las densidades inicial, conjunta bivariante y de transición, $p(\cdot), q(\cdot, \cdot), t(\cdot/x)$, $x \in \mathcal{R}$, respectivamente. Sea \mathbf{K} una f.d.p., con las primeras $n+1$ variables $X_j, j = 1, \dots, n+1$ del proceso, se definen las variables aleatorias

$p_n(x), x \in \mathcal{R}, q_n(y), y \in \mathcal{R}^2$ por las relaciones siguientes:

$$p_n(x) = (nh)^{-1} \sum_{j=1}^n \mathbf{K}\left(\frac{x-X_j}{h}\right) \quad (2.3)$$

$$q_n(y) = q_n(x, x') = (nh)^{-1} \sum_{j=1}^n \mathbf{K}\left(\frac{x-X_j}{\sqrt{h}}\right) \mathbf{K}\left(\frac{x'-X_{j+1}}{\sqrt{h}}\right) \quad (2.4)$$

donde $h = h(n)$ es una sucesión de constantes positivas que satisfacen ciertas condiciones adicionales. Se tiene además el conjunto

$$t_n(x'/x) = \frac{q_n(x', x)}{p_n(x)} \quad (2.5)$$

A partir de $p_n(x)$ y $q_n(y)$ se definen las variables aleatorias,

$$F_n(x) = \int_{-\infty}^x p_n(z) dz, \quad G_n(z/x) = \int_{-\infty}^z t_n(dx'/x)$$

Finalmente sean $F(\cdot)$ y $G(\cdot/x)$ las funciones de distribución inicial y de transición del proceso. Bajo condiciones apropiadas de la función \mathbf{K} , la sucesión $\{h(n)\}$, y el proceso, los resultados son los siguientes:

(i) La función de distribución, $F_n(\cdot)$, como un estimador de $F(x)$ cumple con el Teorema de Glivenko-Cantelli.

(ii) El estimador $G_n(\cdot/x)$ de $G(\cdot/x)$, cumple con el resultado siguiente:

$$\text{SUP}\{|G_n(z/x) - G(z/x)|, z \in \mathcal{R}\}$$

converge en probabilidad a cero, cuando $n \rightarrow \infty, \forall x \in \mathcal{R}$. //

(Yakowitz, 1985) Dada una sucesión (estacionaria) de Markov $\{X_j\}$, la cual presumimos que es G_2 y que tiene f.d.p. continua (estacionaria) $\pi(x)$ y f.d.p. de transición $f(y/x)$. Se hace la siguiente proposición para la estimación de $f(y/x)$.

Proposición: Para las variables aleatorias X_i , la sucesión $\{a(n)\}$, una sucesión de números positivos, y una f.d.p. d -variante $\mathbf{K}(\cdot)$, la expresión siguiente es mejor conocida como un estimador no paramétrico para la densidad común de las X_i 's,

$$P_n(x) = \left[a(n)^d \right]^{-1} \sum_{i=1}^n \mathbf{K}\left(\frac{x-X_i}{a(n)}\right) \quad (2.6)$$

El fundamento de los métodos para hacer inferencias y regresión acerca de la f.d.p. de transición en esta sección y en las siguientes están motivados por el descubrimiento de Rosenblatt (1970) de que bajo el supuesto G_2 , la propiedad de normalidad asintótica de $P_n(x)$ incluso para sucesiones de Markov se cumple. Específicamente, asumimos que los X_i 's son vectores con d coordenadas, que $\mathbf{K}(\cdot)$ es una f.d.p. d -variante que es uniformemente continua la cual satisface

$$\mathbf{K}(u) = O(|u|^{-1}) \text{ cuando } |u| \rightarrow \infty \quad (2.7)$$

$$\int u_j \mathbf{K}(u_1, \dots, u_d) du_1 \dots du_d = 0, \quad 1 \leq j \leq d \quad (2.8)$$

Con estos supuestos se tienen los resultados fundamentales siguientes,

Lemma Sea $\mathbf{K}(\cdot)$ una f.d.p. d -variante, y sea $\{X_n\}$ una sucesión de Markov d -variante con f.d.p. inicial y de transición (estacionarias) $\pi(x)$ y $f(y/x)$ respectivamente, las cuales son acotadas y continuamente diferenciables. Supóngase además que todos los momentos de segundo orden $m_{ij} = \int u_i u_j \mathbf{K}(u) du$ son finitos y que $P_n(x)$ se calcula de acuerdo con (2.6), con $n[a(n)^d] \rightarrow \infty$, pero que $a(n) \rightarrow 0$,

(1) El sesgo

$$E[P_n(x)] - \pi(x) = \frac{1}{2} \sum \left[m_{ij} \frac{d^2}{dx_i dx_j} \pi(x) \right] a(n)^2 + O[a(n)^2] \quad (2.9)$$

(2) Si $\mathbf{K}(x - X_j)$ tiene momento finito de orden cuatro, entonces

$$\left[n a(n)^d \right]^{\frac{1}{2}} \left\{ P_n(z_i) - E[P_n(z_i)] \right\} \sim NID(0, \sigma_i^2) \quad (2.10)$$

$i=1, 2, \dots$; los z_i 's son distintos y

$$\sigma_i^2 = \pi(z_i) \int \mathbf{K}^2(u) du, \quad i=1, \dots, m \quad (2.11)$$

Si $f(y/x) = \frac{f(x,y)}{\pi(x)}$ es la f.d.p. estacionaria para (X_i, X_{i+1}) , entonces, un estimador lógico para la densidad de transición es, por lo tanto,

$$\hat{f}_n(y/x) = \frac{P_n(x,y)}{P_n(x)} \quad (2.12)$$

donde

$$P_n(x,y) = [a(n)^{2d_n}]^{-1} \sum_{j=1}^{n-1} h\left[\frac{(x-X_{i,j}, y-X_{i+1})}{a(n)}\right] \quad (2.13)$$

y $h(z)$ es una f.d.p. bivariante.

Teorema 2.1 Sean $\mathbf{K}(u)$ y $h(z)$ f.d.p.'s en \mathbb{R}^d y \mathbb{R}^{2d} , respectivamente, las cuales satisfacen (2.7) y (2.8) y supóngase que $\{X_n\}$ es una sucesión (estacionaria) de Markov d -variante con f.d.p.'s dos veces diferenciable $\pi(x)$ y $f(x,y)$ para $\{X_i\}$ y $\{(X_i, X_{i+1})\}$ respectivamente. Supóngase además que $na(n)^{2d} \rightarrow \infty$, $\mathbf{K}(x - X_i)$ y $h(x - X_i, y - X_{i+1})$ tienen momentos de cuarto orden, y $na(n)^{2d+4} \rightarrow 0$.

Entonces para $\hat{f}(y/x)$ determinado por (2.12), se tiene que

$$\sqrt{a(n)^{2d_n}} [\hat{f}_n(y/x) - f(y/x)] \quad (2.14)$$

es asintóticamente distribuido normalmente con media cero y varianza igual

$$a \quad V = f(x,y) \frac{\int h^2(z) dz}{[\pi(x)]^2} \quad (2.15)$$

(Hernández-Lerma et al, 1988) Sea $\{X_t : t = 0, 1, \dots\}$ un proceso de markov \mathbb{R}^d -valorado con densidad de transición $q(y|x)$, una distribución inicial arbitraria μ_0 . De donde, para todo $t = 1, 2, \dots$, la distribución de μ_t de X_t está dada recursivamente por

$$\mu_t(B) = \int Q(B|x)\mu_{t-1}(dx), \quad B \in \beta^d \quad (2.16)$$

donde β es el sigma-álgebra de Borel en \mathbb{R}^d , $Q(B|x)$ denota la función de probabilidad de transición, es decir,

$$Q(B|x) = \int_B q(y|x)dy, \quad x \in \mathbb{R}^d \text{ y } B \in \beta^d.$$

Supuestos:

(i) Existe un número positivo $\alpha < 1$ tal que,

$$\|Q(\cdot|x) - Q(\cdot|y)\| \leq 2\alpha, \quad \forall x, y \in \mathbb{R}^d$$

es decir, independientemente de la distribución inicial μ_0 , el proceso de markov tiene distribución invariante μ , esto es,

$$\mu(B) = \int Q(B|x)\mu(dx), \quad \forall B \in \beta^d.$$

Para poder estimar μ , asumimos lo siguiente:

(a) La distribución inicial μ_0 es absolutamente continua con densidad acotada Φ_0 .

(b) Existe una constante q y una función $g \in B(\mathbb{R}^d)$ - espacio de Banach - y $|q(y/x) - q(y'/x)| \leq |g(x)||y - y'| \quad \forall x, y, y' \in \mathbb{R}^d$. De donde, μ_0 es absolutamente continua con densidad Φ_t dada recursivamente por

$$\Phi_t = \int_B q(y/x)\Phi_{t-1}(x)dx, \quad \text{para casi todo } y \in \mathbb{R}^d, \quad t \geq 1$$

como μ es absolutamente continua, existe una función medible no negativa Φ tal que $\mu(B) = \int_B \Phi(x)dx$ para $B \in \beta^d$.

Así, $\Phi(y) = \int q(y/x)\Phi(x)dx$, para casi todo $y \in \mathbb{R}^d$.

(i) Estimación de la Densidad Invariante

Para $n = 0, 1, 2, \dots$. Sea U_n la función definida por $U_n(x) = b_n^d U\left(\frac{x}{b_n}\right)$, $x \in \mathbb{R}^d$ donde U es una f.d.p., algunas veces mencionada como función kernel, y $\{b_n\}$ es una sucesión de números positivos. El estimador recursivo *Wolerton-Wagner (WW)* $\hat{\Phi}_t$ de la densidad es entonces definido como

$$\hat{\Phi}_t(x) = t^{-1} \sum_{n=0}^{t-1} U_n(x_n - x), x \in \mathbb{R}^d, \text{ y } t=1, 2, \dots \quad (2.17)$$

y por otra parte $\mu_t(B) = \int_B \hat{\Phi}_t(x) dx$, se define como un estimador para la función de probabilidad invariante μ . Asumiendo que el kernel U está acotado, $\|U\| < \infty$ y también que $\mu = \int |x| U(x) dx < \infty$.

Con respecto a la sucesión $\{b_t\}$, asumimos que es no creciente, y que satisface alguna de las condiciones siguientes:

Condición 1

- (a) $b_t \rightarrow 0$
- (b) $t b_t^d \rightarrow \infty$
- (c) $\sum_t t^{-\frac{3}{2}} b_t^{-d} < \infty$

Para establecer nuestro primer resultado consistente, definimos la función sesgo, $B_t(x) = E[\hat{\Phi}_t(x)] - \Phi(x)$ y el error cuadrático medio,

$$M_t(x) = E[\hat{\Phi}_t(x) - \Phi(x)]^2 = V[\hat{\Phi}_t(x)] + B_t(x)^2$$

Teorema 2.2 Supóngase que la condición 1(a) se cumple, entonces para $t \rightarrow \infty$, se tiene que

- (a) $SUP_x |B_t(x)| \rightarrow 0$

Además si la condición 1(b) se cumple, entonces

- (b) $SUP_x Var[\hat{\Phi}_t(x)] \leq ct^{-1} b_t \rightarrow 0$, donde $c = 2\|U\| \frac{C_o}{(1-\alpha)}$, y

- (c) $SUP_x M_t(x) \rightarrow 0$

Supóngase también que la condición 1(c) se cumple, entonces,

- (d) $\hat{\Phi}_t(x) \rightarrow \Phi(x)$, casi seguramente, $\forall x \in \Re^d, y$
(e) $\|\mu_t - \mu\| = \int |\hat{\Phi}_t(x) - \Phi(x)| dx$ casi seguramente.

(ii) Estimación de la Densidad Conjunta

Sea $Z_t = (X_t, X_{t+1})$, $t = 0, 1, \dots$, el cual es un proceso de markov bidimensional, y $\{Z_t\}$ tiene densidad $\hat{f}(x, y) = q(y/x)\hat{\Phi}_t(x)$, $t \geq 0$, la cual converge en norma L_1 a $f(x, y) = q(y/x)\Phi(x)$.

Para definir un estimador $\hat{f}_t(x, y)$ de la densidad conjunta $f(x, y)$, consideremos de nuevo la sucesión $\{b_n\}$ y las funciones $U(x)$ y $U_n(x)$ introducidas en la sección anterior, y sea

$$U_n^*(x, y) = U_n(x)U_n(y) = b_n^{-2d} U\left(\frac{x}{b_n}\right)U\left(\frac{y}{b_n}\right), \quad \forall (x, y) \in \Re^{2d},$$

luego

$$\hat{f}(x, y) = t^{-1} \sum_{n=0}^{t-1} U_n^*(x_n - x, x_{n+1} - y) \quad (2.18)$$

Para establecer el correspondiente resultado consistente, introducimos la función sesgo, $B_t^*(x, y) = E[\hat{f}_t(x, y)] - f(x, y)$ y la función error cuadrático medio

$$M_t^*(x, y) = [E[\hat{f}_t(x, y)] - f(x, y)]^2 = V[\hat{f}_t(x, y)] - B_t^*(x, y)$$

Considerando las condiciones 1(b-c) y,

Condición 2

- (a) $tb_t^{2d} \rightarrow \infty$
(b) $\sum_t t^{-\frac{3}{2}} b_t^{-2d} < \infty$

Teorema 2.3 Supóngase que la condición 1(a) se tiene, esto es, $b_t \rightarrow 0$ cuando $t \rightarrow \infty$, se cumple que:

- (a) $SUP_{x,y} |B_t^*(x, y)| \rightarrow 0$
si en adición, la condición 2(a) se cumple, entonces:

- (b) $SUP_{x,y} \{ V[\hat{\mathbf{f}}_t(x,y)] \} < ctb_t^{-2d}$, para alguna constante c , y
(c) $SUP_{x,y} M_t^*(x,y) \rightarrow 0$.

Supóngase también que la condición 2(b) se cumple. Entonces,

- (d) $\hat{\mathbf{f}}_t(x,y) \rightarrow f(x,y)$ casi seguramente, $\forall (x,y) \in \mathbb{R}^{2d}$, y
(e) $\iint |\hat{\mathbf{f}}_t(x,y) - f(x,y)| dx dy \rightarrow 0$ casi seguramente.

(iii) Estimación de la Densidad de Transición

Teniendo dos estimadores $\hat{\Phi}_t(x)$ y $\hat{\mathbf{f}}_t(x,y)$ de $\Phi(x)$ y $f(x,y)$, respectivamente, ahora podemos definir un estimador $\hat{q}_t(y/x)$ para la densidad de transición $q(y/x) = \frac{f(x,y)}{\Phi(x)}$ en una forma obvia:

$$\hat{q}_t(y/x) = \frac{\hat{\mathbf{f}}_t(x,y)}{\hat{\Phi}_t(x)}, \quad \forall (x,y) \in \mathbb{R}^{2d}, t \geq 1 \quad (2.19)$$

Teorema 2.4 Sea $x \in \mathbb{R}^d$ tal que $\Phi(x) > 0$. Si los supuestos de los teoremas 2.2(d) y 2.3(c) se cumplen, entonces, cuando $t \rightarrow \infty$, se tiene que

- (a) $SUP_y E[\hat{q}(y/x) - q(y/x)]^2 \rightarrow 0$

Si también la condición 2(b) se cumple, entonces,

- (b) $\hat{q}_t(y/x) \rightarrow q(y/x)$ casi seguramente $\forall y \in \mathbb{R}^d$, y
(c) $\left\| \hat{Q}_t(\cdot/x) - Q(\cdot/x) \right\| = \int |\hat{q}_t(y/x) - q(y/x)| dy \rightarrow 0$ casi seguramente,
de donde,

$$\hat{Q}_t(B/x) = \int_B \hat{q}_t(y/x) dy, \quad B \in \beta^d \quad (2.20)$$

es un estimador de la función de probabilidad de transición $Q(B/x)$. //

CAPITULO III

3.- METODOLOGIA ESTADISTICA

3.1.- METODO BOOTSTRAP

3.1.1.- Introducción

Un problema típico en Estadística generalmente involucra la estimación de un parámetro desconocido . Las principales preguntas a responder inicialmente son (1) ¿Cuál estimador $\hat{\Theta}$ debería ser utilizado?; y (2) ¿Qué tan preciso es como estimador de θ ?

Antes de comenzar a describir cómo trabaja el bootstrap, vamos a describir una situación donde éste no es necesario. Supóngase que nuestros datos consisten de una muestra aleatoria de tamaño n , X_1, \dots, X_n de una función de distribución F (desconocida) y queremos estimar un funcional, tal como la media poblacional, μ , $\mu = \int x dF(x)$ consideremos entonces utilizar el mismo funcional, pero de la función de distribución muestral, F_n , el cual en su lugar es la media muestral, esto es, $\bar{X} = \int x dF_n(x)$, y queremos conocer la precisión de la media muestral, \bar{X} , como estimador de μ , la

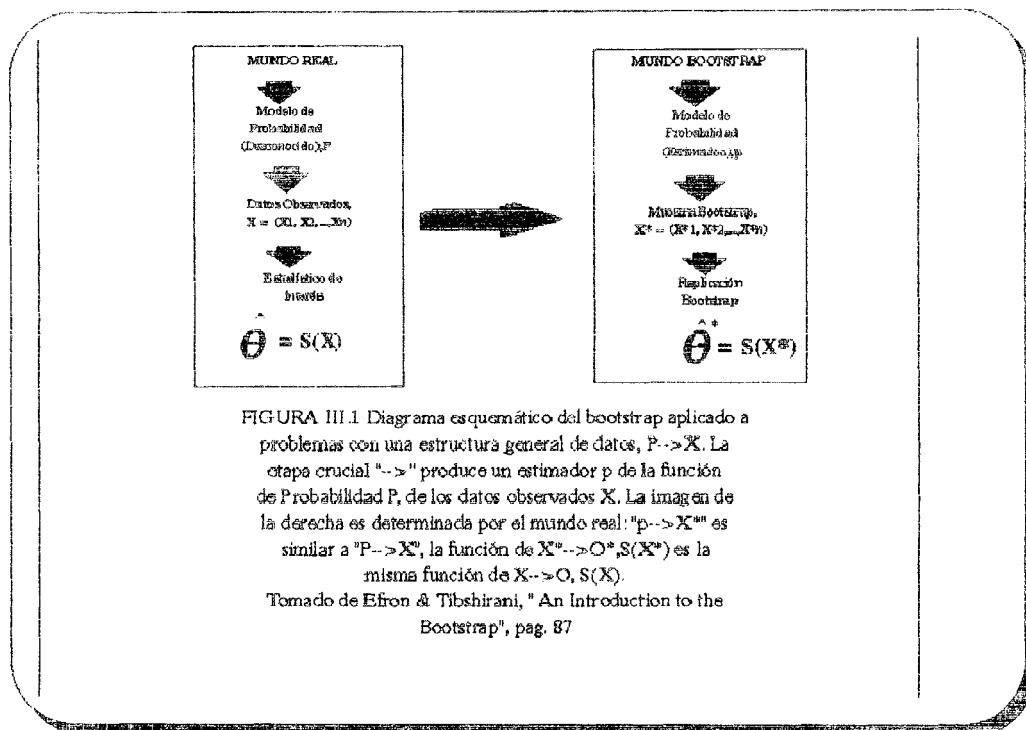
media poblacional. Si el segundo momento de F con respecto a μ es $\mu_2 = E_F(X^2) - \mu_X^2$, entonces el error estándar, $\sigma(F, n, \bar{X})$, es decir, la desviación estándar \bar{X} para una m.a.n. de la distribución F es $\sigma(F) = \left[\frac{\mu_2(F)}{n} \right]^{\frac{1}{2}}$. El error estándar es la medida de precisión tradicional para la media muestral. Desafortunadamente no podemos utilizar $\sigma(F)$ ya que no conocemos $\mu_2(F)$, pero, si podemos utilizar el estimador del error estándar, $se = \left[\frac{S_2}{n} \right]^{\frac{1}{2}}$, donde S_2 es la desviación estándar muestral. Existe una forma más obvia para estimar $\sigma(F)$.

Sea F_n la función de distribución muestral, entonces el estimador del error estándar de la media muestral $\sigma(\cdot)$ nos queda de la forma siguiente $\sigma(F_n) = \sqrt{\frac{\mu_2(F_n)}{n}}$. Este procedimiento se denomina el *Principio Plug-in*, el cual es un método simple de estimación de parámetros a partir de una muestra aleatoria. El estimador plug-in de un parámetro $\theta = t(F)$ está definido por $\hat{\theta} = t(F_n)$. Es decir vamos a estimar la función $\theta = t(F)$ de la función de distribución F , por la misma función, pero de la función de distribución empírica F_n , $\hat{\theta} = t(F_n)$. El método bootstrap es en si mismo una aplicación de este principio.

Este argumento no siempre es practicable - una función de densidad de probabilidad es sólo uno de los ejemplos de un funcional de F que no es directamente manejable con este tratamiento. Sin embargo, *Efron* en 1979, introdujo el *Método Bootstrap*, el cual consiste en remuestrear las observaciones muestrales previamente seleccionadas de una manera particular con el objetivo de estimar parámetros poblacionales.

La esencia de éste método es la simulación de propiedades de un procedimiento estadístico con un mínimo de suposiciones. La palabra

simulación se utiliza en un amplio sentido: desde una simple sustitución de una distribución estimada en una fórmula, hasta simulación de Monte Carlo de muestras aleatorias y sus respectivos análisis.



El método bootstrap es una técnica de reciente desarrollo para hacer ciertos tipos de inferencias estadísticas, la cual requiere de la capacidad y velocidad de los modernos computadores, para poder simplificar los complicados cálculos que con frecuencia aparecen en la Teoría Estadística tradicional. El objetivo particular de la teoría bootstrap es una implementación muy particular de los conceptos estadísticos en base a computación intensiva, los cuales son más sencillos de comprender desde un contexto computacional.

que a través de la exposición matemática usual. En cierto modo, el nombre *Bootstrap* transmite la impresión a los estadísticos de *sacar algo de la nada*, los cuales idealmente remuestrean de sus muestras, que presumiblemente tendrían tanto éxito como si ellos trataran de levantarse por los cordones de sus zapatos, sin embargo, el bootstrap es una técnica de base teórica promisoria.

Ahora vamos a introducir alguna notación, una muestra $X' = (X_1, \dots, X_n)$ es una colección de n números (no necesariamente escalares), sin considerar el orden, seleccionadas aleatoriamente de la población con función de distribución F . Por aleatoriedad entendemos que los X_i' s son variables aleatorias independientes e idénticamente distribuidas cada una con función de distribución F .

En los denominados problemas no paramétricos una muestra $X^{*'} = (X_1^*, \dots, X_n^*)$ es una colección no ordenada de ítems seleccionados aleatoriamente de X , de modo que cada X_i^* tiene igual probabilidad de ser igual a cualquiera de los X_j 's,

$$P(X_i^* = X_j / X) = n^{-1}, 1 \leq i, j \leq n.$$

Es decir, los X_i^* 's son variables independientes e idénticamente distribuidas, condicionadas por X , con esta distribución. Por supuesto, esto significa que X^* es probable que contenga valores repetidos, los cuales todos serán listados en la colección X^* . En problemas paramétricos X^* denota una muestra seleccionada aleatoriamente de una población dependiente de un parámetro cuyos valores han sido estimados. Si la población es continua entonces con probabilidad uno, todos los valores en X^* son diferentes. En ambos casos problemas paramétricos y no

paramétricos, F_n denota la función de distribución de la "población" de la cual X^* fue seleccionada.

Un estimador $\hat{\theta}$ es una función de los datos, y puede también considerarse como un funcional de la función de distribución empírica, F_n , aunque son funciones numéricamente equivalentes, es útil distinguir entre ellas, para ello utilizaremos corchetes para la primeras y paréntesis para la segunda, $\hat{\theta} = \theta[X] = \theta(F_n)$. Por ejemplo, si $\theta = \theta(F) = \int x dF(x)$, es la media poblacional y F_n es la función de distribución empírica (la cual asigna una ponderación de n^{-1} a cada punto X_i), entonces,

$$\begin{aligned}\hat{\theta} &= \theta[X] = \frac{1}{n} \sum_{i=1}^n X_i \\ &= \theta(F_n) = \int x dF_n(x)\end{aligned}$$

la media muestral.

3.1.2.- Método Bootstrap

Sea $X' = (X_1, \dots, X_n)$ una muestra aleatoria proveniente de una distribución desconocida F , sea $x' = (x_1, \dots, x_n)$ la muestra aleatoria observada. Sea $\theta(F)$ algún parámetro poblacional de interés, tal como la media de F , y sea $t(X)$ un estimador de $\theta(F)$ tal como la media muestral.

Sea,

$$H(F, X) = t(X) - \theta(F) \quad (3.1)$$

Supongamos que estamos interesados en estimar a partir de los datos observados x , la distribución muestral de H , o algún otro parámetro de interés como la media de H - el sesgo de $t(X)$ -, la varianza de H - la varianza de $t(X)$, para ello aplicaremos el siguiente algoritmo.

3.1.2.1.- Algoritmo del Método Bootstrap

(i) Construimos la función de distribución empírica de los datos,

$$F_n: \text{masa de probabilidad } n^{-1} \text{ para } x_i, i = 1, \dots, n \quad (3.2)$$

(ii) Con F_n fija seleccionamos una muestra aleatoria de tamaño n de F_n ,

$$X_i^* = x_i^*, X_i^* \sim IND F_n, i = 1, 2, \dots, n \quad (3.3)$$

Esta muestra es la denominada **muestra bootstrap**.

(iii) Aproximar la distribución muestral de $H(X, F)$ por la distribución bootstrap de

$$H^*(X^*, F_n) \quad (3.4)$$

es decir, la distribución de H^* inducida por el mecanismo aleatorio (3.3) con F_n fija.

La dificultad del método bootstrap está en el cálculo de la distribución bootstrap, la cual es en la mayoría de los casos imposible de calcular, sin embargo existen tres métodos que hacen posible este cálculo:

(1) El cálculo directo;

(2) Aproximación de la Distribución Bootstrap por el método de Monte Carlo: repetidas realizaciones de X^* generadas a partir de la muestra observada de tamaño n provenientes de F_n , digamos X_1^*, \dots, X_B^* ; y el histograma de los correspondientes valores $H(X_1^*), \dots, H(X_B^*)$ se toma como una aproximación de la distribución bootstrap de $H(X^*, F_n)$, la cual mejora a medida que el número de muestras se incrementa.

Este enfoque funciona, porque en la mayoría de las aplicaciones, la cantidad bootstrap puede expresarse como una esperanza condicional sobre la muestra o equivalentemente como una integral con respecto a la función de

distribución empírica. Nuestro desarrollo de los Principios del Bootstrap, se enfoca mayormente sobre esta propiedad.

(3) Utilizar Series de Taylor para aproximar la media y la varianza de la distribución bootstrap de H^* .

3.1.3.- Errores Estándares y Estimadores de Errores Estándares

3.1.3.1.- Introducción

Los estadísticos tales como $\hat{\theta} = t(F_n)$ con frecuencia son los primeros resultados de un análisis de datos, lo siguiente que se tiene que hacer es conocer la precisión de éste, el bootstrap provee estimadores precisos para estimar el error estándar del estadístico de interés mediante el principio plug-in. Supóngase que se tiene una muestra aleatoria $X = (X_1, \dots, X_n)$ de una distribución de probabilidad F (desconocida) la cual ha sido observada y queremos estimar un parámetro de interés $\theta = t(F)$ con base en X . Para este propósito, calculamos el valor de un estadístico $\hat{\theta} = S(X)$ de X . ¿Qué tan preciso es $\hat{\theta}$?

3.1.3.2.- El Estimador Bootstrap del Error Estándar

Sea F_n la función de distribución empírica, asignando probabilidad n^{-1} sobre cada valor X_i , $i=1, \dots, n$, observado. Una muestra bootstrap es una muestra aleatoria de tamaño n seleccionada de F_n , digamos $X^* = (X_1^*, \dots, X_n^*)$, es decir,

$$X_1^*, \dots, X_n^* \sim \text{IND}F_n \quad (3.5)$$

Existe otra forma de decirlo: las observaciones X_j^* , $j=1, \dots, n$ son una muestra aleatoria seleccionada con reposición de una población con n objetos (X_1, \dots, X_n) .

De esta forma, podríamos tener $X_1^* = X_7, X_2^* = X_7, \dots, X_n^* = X_{n-1}$. El conjunto de datos X^* consiste de miembros del conjunto original (X_1, \dots, X_n) , algunos apareciendo cero veces, algunos apareciendo una vez, algunos apareciendo dos veces, etc.

Dado un conjunto de datos $X^*, \hat{\theta}^* = S(X^*)$ se denomina una replicación bootstrap de $\hat{\theta}$. Es decir, la cantidad $S(X^*)$ es el resultado de aplicar la misma función $S(\cdot)$ a X^* tal como fue aplicada a X . Por ejemplo, si $S(X)$ es la media muestral, entonces $S(X^*)$ es la media muestral de los datos bootstrap.

El estimador bootstrap de $se_F(\hat{\theta})$, el error estándar de $\hat{\theta}$, es un estimador plug-in que utiliza la función de distribución empírica (F.D.E.) en lugar de la F.D. F . Específicamente, el estimador bootstrap de $se_F(\hat{\theta})$ es $se_{F_n}(\hat{\theta}^*)$, es decir, el error estándar de $\hat{\theta}^*$ para conjuntos de datos de tamaño n seleccionados aleatoriamente de F_n , también denominado Estimador Bootstrap ideal del error estándar de $\hat{\theta}$.

El algoritmo bootstrap, próximo a describir es una forma computacional de obtener una buena aproximación del valor numérico de $se_{F_n}(\hat{\theta}^*)$, y es fácil de implementar en el computador.

Algoritmo del Método Bootstrap para estimar Errores Estándares

1.- Seleccionar B muestras independientes bootstrap X_1^*, \dots, X_B^* , cada una consistente de n valores seleccionados aleatoriamente con reposición de X , para estimar un error estándar, el número B generalmente estará entre 25 y 200.

2.- Evaluar la replicación $\hat{\theta}^*(b) = S(X_b^*)$, $b = 1, \dots, B$, en cada una de las muestras bootstrap.

3.- Estimar el error estándar $se_F(\hat{\theta})$, por la desviación estándar de la B replicaciones $\hat{\theta}^*$,

$$se_B = \sqrt{\frac{\sum_{b=1}^B [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1}},$$

donde

$$\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^B \hat{\theta}^*(b)}{B}.$$

El algoritmo bootstrap trabaja seleccionando un número finito de muestras bootstrap, evaluando las correspondientes replicaciones y estimando el error estándar de $\hat{\theta}$, por la desviación estándar de las replicaciones. El resultado se denomina el error estándar denotado por se_B , donde B es el número de muestras bootstrap utilizadas. El estimador bootstrap ideal $se_F(\hat{\theta}^*)$ y su aproximación se_B algunas veces se denominan estimadores bootstrap no paramétricos porque están basados en F_n , el estimador no paramétrico de F .

3.1.3.3.- El Número de Replicaciones Bootstrap B

¿Qué tan grande debería ser B , el número de replicaciones bootstrap utilizadas para evaluar se_B ? La cantidad de tiempo que toma evaluar las replicaciones bootstrap se incrementa linealmente con B . Las restricciones de tiempo pueden sugerir un valor pequeño de B si la función $S(X)$ es muy complicada. Existen dos reglas, obtenidas a partir de la experiencia:

- (i) Un número de replicaciones pequeño, digamos $B = 25$, usualmente es informativo; $B = 50$ con frecuencia es suficiente para dar un buen estimador del error estándar.

(ii) Muy raras veces son necesarias más de $B = 200$ replicaciones para estimar un error estándar (valores más grandes de B son necesarios para Intervalos Confidenciales Bootstrap).

La gran ventaja del bootstrap es que pueden calcularse tantas replicaciones $\hat{\theta}^*$ como se desee, o al menos tantas replicaciones como se permitan. Esto nos permite efectuar cálculos probabilísticos directamente, por ejemplo, utilizando la variabilidad observada de los $\hat{\theta}^*$'s para estimar la cantidad no observable $se_F(\hat{\theta})$.

El Método Bootstrap tiene en cierto modo ventajas diferentes sobre los métodos encontrados en los textos tradicionales: (i) Cuando se utiliza de modo no paramétrico, éste releva al analista de tener que hacer supuestos paramétricos acerca de la forma de la población fundamental; (ii) Cuando se utiliza de modo paramétrico, provee respuestas más precisas que las fórmulas encontradas en los textos y respuestas en problemas para los cuales no existen fórmulas.

3.1.3.4.- Estructuras de Datos más Complicadas

El algoritmo bootstrap estudiado anteriormente está basado en el modelo probabilístico más sencillo: el modelo univariante, donde sólo una distribución de probabilidad, F (desconocida) produce los datos X por muestreo aleatorio, es decir, $X_1, \dots, X_n \sim \text{IND}F$. Los datos X_i pueden ser más complicados siendo quizás escalares o vectores o funciones pero la función de probabilidad es simple. En Estadística, algunos análisis de datos involucran estructuras de datos más complicadas tales como series de tiempo, análisis de varianza, modelos de regresión, etc.

El algoritmo bootstrap puede adaptarse para estas estructuras de datos y otras más generales. La mayor parte de la literatura del método bootstrap concentra mayormente su atención en observaciones independientes. Sin embargo, estructuras de datos, tales como series de tiempo, procesos de markov, involucran observaciones no independientes. ¿ Puede aplicarse el método bootstrap sobre observaciones no independientes?, por ejemplo, supóngase que se tiene el proceso autoregresivo de primer orden, $X_t = \beta X_{t-1} + \xi_t$, $\xi_t \sim NID(0, \sigma^2)$, y $X_0 = 0$. Si seleccionamos una muestra aleatoria con reposición $\{X_t^*: t = 1, \dots, N\}$ de los X_t s, se perdería esta relación de no independencia existente entre X_{t-1} y X_t , debido al mecanismo de selección. Existen formas alternas que tratan de conservar esta relación de dependencia entre los X_t s al construir la muestra bootstrap.

Alternativa 1 Sea $X_t = \beta X_{t-1} + \xi_t$, calcular los errores estimados $\hat{\xi}_t = X_t - \hat{\beta}X_{t-1}$, donde $\hat{\beta}$ se obtiene a partir de los X_t s originales, calcular $\tilde{\xi}_t = \hat{\xi}_t - n^{-1} \sum_{t=1}^n \hat{\xi}_t$, los errores centrados. Sea \tilde{F}_n la función de distribución empírica de los $\tilde{\xi}_t$ s. Pretendiendo que \tilde{F}_n es la verdadera distribución seleccionamos una muestra aleatoria $\{\tilde{\xi}_t^*, t = 1, \dots, N\}$ de \tilde{F}_n . De este modo las variables aleatorias $\{\tilde{\xi}_t^*, t = 1, \dots, N\}$ son independientes e idénticamente distribuidas con función de distribución \tilde{F}_n . Ahora construimos la muestra bootstrap $\{X_t^*, t = 1, \dots, N\}$ recursivamente por $X_t^* = \hat{\beta}X_{t-1}^* + \tilde{\xi}_t^*$, $t = 1, \dots, N$ y $X_0^* = 0$.

Alternativa 2 Esta es una alternativa diferente de aplicar el método bootstrap en series de tiempo. En lugar de remuestrear los residuos, se obtienen "bloques" de observaciones consecutivas, por ejemplo, si la serie

es X_1, X_2, \dots, X_6 y seleccionamos bloques de longitud 3, entonces tendríamos los bloques:

$$(X_1, X_2, X_3), (X_2, X_3, X_4), (X_3, X_4, X_5), (X_4, X_5, X_6).$$

entonces muestreamos los bloques con reposición, y luego se "pegan" para formar la serie de tiempo bootstrap. Se seleccionan tantos bloques como sea necesario para obtener una serie de aproximadamente la misma longitud que la serie original. Si la longitud del bloque es l , entonces seleccionamos k bloques de modo tal que $n \approx k \cdot l$. La justificación para este procedimiento es simple, ya que no podemos remuestrear las observaciones individuales, porque destruiríamos la relación de no independencia que estamos tratando de capturar. Con el procedimiento antes mencionado, la idea es seleccionar un bloque de longitud l suficientemente grande de modo que las observaciones que se encuentran fuera del bloque sean independientes, mientras que las observaciones integrantes del bloque retienen la relación de no independencia. El mayor inconveniente de este procedimiento es que la serie de tiempo original debe tener una longitud considerable.

Alternativa 3 Supóngase que se tienen las observaciones X_i , $i = 1, \dots, n$, en donde la relación de no independencia entre X_{t-1} y X_t puede ser lineal o no lineal. Además $(X_{t-1}, X_t) \sim f_{X_{t-1}, X_t}(x_{t-1}, x_t)$, la cual es desconocida, y queremos realizar inferencias acerca de algún estadístico de nuestro interés que relaciona las variables aleatorias X_{t-1} y X_t . Entonces un mecanismo de remuestreo que trata la relación de no independencia de las observaciones X_t 's es el siguiente: Construir los pares consecutivos $(X_1, X_2), \dots, (X_{n-1}, X_n)$, en total $n-1$ pares. Tomando estos pares

como una muestra aleatoria de $f_{X_{t-1}, X_t}(\cdot, \cdot)$, construir la función de distribución empírica, F_n , de estas observaciones (pares) y seleccionar una muestra aleatoria, con reposición, $\left\{ (X_{1i}^*, X_{2j}^*) : i, j = 1, \dots, n-1 \right\}$. Luego utilizar la serie X_{1i}^* o X_{2j}^* como muestra bootstrap univariante.

3.1.4.- Estimadores del Sesgo

El sesgo de $\hat{\theta} = S(X)$ como un estimador de θ está definido como la diferencia entre el valor esperado de $\hat{\theta}$ y el valor del parámetro θ ,

$$Sesgo_F = Sesgo_F(\hat{\theta}, \theta) = E_F[S(X)] - t(F) \quad (3.6)$$

Un valor grande del sesgo es un aspecto indeseable en el desarrollo de los estimadores. Los estimadores insesgados aquellos para los cuales $E_F[\hat{\theta}] = \theta$, tienen un rol importante en la teoría estadística y en la práctica. Estos producen un sentimiento de objetividad científica en los procesos de estimación. Podemos utilizar el bootstrap para asignar el sesgo de cualquier estimador, el estimador bootstrap del sesgo está definido como $sesgo_{F_n}$, sustituyendo F_n por F , en (3.6),

$$Sesgo_{F_n} = E_{F_n}[S(X^*)] - t(F_n) \quad (3.7)$$

donde $E_{F_n}[S(X^*)]$ puede aproximarse por el promedio de las replicaciones $\hat{\theta}_i^*$, $i = 1, \dots, B$, esto es,

$$\hat{\theta}^*(\cdot) = \frac{\sum_{i=1}^B \hat{\theta}_i^*}{B} = \frac{\sum_{i=1}^B S(X_i^*)}{B} \quad (3.8)$$

El estimador bootstrap del sesgo basado en B replicaciones bootstrap, $sesgo_B$ con $\hat{\theta}^*(\cdot)$ sustituido por $E_{F_n}[S(X^*)]$, es

$$sesgo_B = \hat{\theta}^*(\cdot) - t(F_n) \quad (3.9)$$

3.1.5.- Propiedades Asintóticas del Método Bootstrap¹

Sea $X = (X_1, \dots, X_n)$ una muestra aleatoria proveniente de una función de distribución F (desconocida). Sea $T_n = T_n(X_1, \dots, X_n)$ un estadístico y $T: D \rightarrow \mathbb{R}$ es un funcional (parámetro), donde D es una clase de funciones que contiene (al menos) a F y a todas las distribuciones discretas con soporte infinito.

Sea $H_n(X, F) = \sqrt{n}[T_n - T(F)]$, y supongamos que la distribución asintótica de $H_n(X, F)$ es normal con media cero.

Dada la muestra observada $x = (x_1, \dots, x_n)$, sea F_n la función de distribución empírica de los datos y con ella obtenemos la muestra bootstrap $X^* = (X_1^*, \dots, X_n^*)$.

Sean $H^* = H_n^*(X^*, F_n) = \sqrt{n}[T_n^* - T(F_n)]$ y f_n^* la distribución de H^* , entonces f_n^* converge, en probabilidad a una distribución normal estándar, si T es una clase particular de funcionales, es decir, $f_n^*(x, F_n)$ converge en probabilidad a la misma distribución.

Caso Particular: La Media Muestral

Sean X_1, \dots, X_n variables aleatorias independientes e idénticamente distribuidas con función de distribución F , sean también μ y σ^2 , la media y la varianza (finitas) de la función de distribución F .

El estimador tradicional de μ es \bar{X} , la media muestral, y su error estándar es $\frac{S_n}{\sqrt{n}}$, donde S_n es la desviación estándar muestral. De acuerdo con el

Teorema Central del Límite, sabemos que la distribución de

$$Q_n = \frac{\sqrt{n}(\bar{X} - \mu)}{S_n} \rightarrow N(0, 1) \quad (3.10)$$

¹ Para una explicación más detallada ver Goitia, 1991, p19-27

converge débilmente a la distribución normal estándar, por tanto, se conoce la distribución asintótica de Q_n .

Lo que ahora se plantea es si a la distribución bootstrap podemos hallarle un comportamiento asintótico similar al de Q_n . En verdad, puede demostrarse que la distribución asintótica de

$$Q_m^* = \frac{\sqrt{m}(\bar{X}^* - \bar{X})}{S_m^*} \rightarrow N(0, 1) \quad (3.11)$$

es normal estándar, donde \bar{X}^* y S_m^* se basan en replicaciones bootstrap.

3.1.6.- Intervalos Confidenciales Bootstrap

Los intervalos confidenciales exactos de un parámetro real θ , de una familia multiparamétrica, en muchos casos no se conocen, solo se conocen aproximaciones.

Supóngase que se tiene una muestra aleatoria de tamaño n , $X = (X_1, \dots, X_n)$ de una distribución desconocida F . Sea $\hat{\theta} = t(F_n)$, el estimador plug-in del parámetro de nuestro interés y sea se alguna estimación razonable del error estándar de $\hat{\theta}$ basado quizás en cálculos jackknife o bootstrap. En la mayoría de los casos, la distribución de $\hat{\theta}$ se approxima a una distribución normal con media θ y desviación estándar se cuando $n \rightarrow \infty$, es decir, $\hat{\theta} \sim N(\theta, se)$ o equivalentemente,

$$Z = \frac{(\hat{\theta} - \theta)}{se} \sim N(0, 1) \quad (3.12)$$

El resultado anterior es asintótico o para muestras grandes, y es verdadero para modelos de probabilidad generales.

Sea $Z^{(\alpha)}$ el α -ésimo percentil de la distribución normal estándar, es decir, si tomamos a (3.12) como verdadero, se tiene que

$$P_F(Z^{(\alpha)} \leq \frac{(\hat{\theta} - \theta)}{se} \leq Z^{(1-\alpha)}) = 1 - 2\alpha \quad (3.13)$$

el cual puede escribirse como:

$$P_F\left[\hat{\theta} - Z^{(1-\alpha)}se \leq \theta \leq \hat{\theta} + Z^{(\alpha)}se\right] = 1 - 2\alpha$$

En general,

$$\left[\hat{\theta} - Z^{(1-\alpha)}se, \hat{\theta} + Z^{(\alpha)}se\right] \quad (3.14)$$

se denomina *Intervalo Confidencial Estándar con probabilidad 1-2α*, o nivel de confianza $100(1-2\alpha)\%$, o más simplemente intervalo confidencial $1-2\alpha$ para θ . Ya que $Z^{(\alpha)} = -Z^{(1-\alpha)}$, (3.14) puede escribirse como, $\hat{\theta} \pm Z^{(1-\alpha)}se$.

3.1.6.1.- El Intervalo *t-bootstrap*

Los intervalos de confianza aproximados basados en cálculos bootstrap fueron introducidos por *Efron* (1981). Igual que los intervalos estándares, estos pueden ser aplicados automáticamente en casi todas las situaciones.

A través del uso del bootstrap podemos obtener intervalos muy precisos sin tener que hacer supuestos de normalidad como en el caso anterior. El enfoque *t-bootstrap* estima directamente la distribución del pivote (3.12), de los datos, es decir, construye la función de distribución empírica de los datos que se tienen a la mano. Este genera B muestras bootstrap y se calcula la versión de (3.12) en cada una de ellas. Más formalmente, el método *t-bootstrap* consiste en:

- (i) Generar B muestras bootstrap: X_1^*, \dots, X_B^*

(ii) Calcular $Z^*(b) = \frac{(\hat{\theta} - \hat{\theta})}{se(\hat{\theta}(b))}$ donde $\hat{\theta}^*(b) = S(X^*b)$ es el valor de $\hat{\theta}$ para la muestra bootstrap X^*b y $se^*(b)$ es el estimador del error estándar de $\hat{\theta}^*$ para la muestra bootstrap X^*b . El α -ésimo percentil de $Z^*(b)$ es estimado por el valor $t^{(\alpha)}$ tal que $\frac{\#\{Z^*(b) \leq t^{(\alpha)}\}}{B} = \alpha$. Por ejemplo, si $B = 1000$, el estimador del punto 5% es el 50^{vo} valor más grande de $Z^*(b)$. Finalmente, el intervalo t -bootstrap esta dado por

$$[\hat{\theta} - t^{(1-\alpha)} se, \hat{\theta} - t^{(\alpha)} se] \quad (3.15)$$

Si $B \cdot \alpha$ no es un entero, el procedimiento siguiente puede ser utilizado. Asumiendo que $\alpha \leq 0.5$, sea $k = [(B+1)\alpha]$ el más grande entero menor o igual que $(B+1)\alpha$. Entonces definimos los cuantiles empíricos α y $1-\alpha$ por el k -ésimo valor mayor y por el $(B+1-k)$ -ésimo valor mayor de $Z^*(b)$, respectivamente.

La cantidad (3.12) se denomina una aproximación pivotal, esto significa que su distribución es aproximadamente la misma para cada valor de θ . Puede demostrarse que para muestras grandes la cobertura del intervalo t -bootstrap tiende a ser más cercana que la cobertura del intervalo confidencial estándar. El intervalo t -bootstrap es particularmente aplicable a estadísticos de localización tales como la media muestral, percentiles muestrales, media trimmed. Pero no puede aplicarse a estadísticos generales como el coeficiente de correlación.

3.1.6.2.- El Intervalo Percentil

Los percentiles del histograma bootstrap pueden utilizarse para definir los límites confidenciales. Supóngase que nos encontramos en la situación general de la fig. III.1. Un conjunto de B datos bootstrap es generado X_1^*, \dots, X_B^* , y las replicaciones bootstrap $\hat{\theta}_b^* = S(X_b^*)$ son calculadas.

Sea G la función de distribución empírica de $\hat{\theta}^*$. El **Intervalo Percentil $1-2\alpha$** está definido por los percentiles α y $1-\alpha$ de G , esto es,

$$[\hat{\theta}^*(\alpha), \hat{\theta}^*(1-\alpha)] = [G^{-1}(\alpha), G^{-1}(1-\alpha)] \quad (3.16)$$

La expresión (3.16) se refiere a la situación bootstrap ideal, en la cual el número de replicaciones es infinito. En la práctica tenemos que utilizar un número finito de replicaciones bootstrap, entonces el intervalo percentil $1-2\alpha$ aproximado es,

$$[\hat{\theta}_B^*(\alpha), \hat{\theta}_B^*(1-\alpha)] = [\hat{G}_B^{-1}(\alpha), \hat{G}_B^{-1}(1-\alpha)] \quad (3.17)$$

donde $\hat{\theta}_B^*(\alpha)$ y $\hat{\theta}_B^*(1-\alpha)$ son los percentiles bootstrap α y $1-\alpha$, respectivamente, de la distribución de $\hat{\theta}_B^*$. Así, si $B = 2000$, y $\alpha = 0.05$, $\hat{\theta}_B^*$, es el 100^{mo} valor ordenado de las replicaciones $\hat{\theta}^*$ (si $B \cdot \alpha$ no es un entero puede utilizarse la convención del intervalo *t-bootstrap*). Si la distribución de $\hat{\theta}^*$ es más o menos normal, entonces el intervalo normal estándar y el intervalo percentil serán más o menos similares.

3.1.6.3.- El Método Bias-Corrected and Accelerated (*BCa*)

Uno de los principales objetivos de la teoría bootstrap es producir "*buenos*" intervalos confidenciales. "*Bueno*" en este contexto significa que los intervalos bootstrap deberán ser cercanos a los intervalos confidenciales exactos en aquellas situaciones donde la teoría estadística suministra una respuesta exacta y debería dar coberturas de probabilidad precisas (exactas). Ninguno de los métodos estudiados cumplen con este criterio: los intervalos *t-bootstrap* tienen buenas coberturas de probabilidad, pero tienden a ser erráticos en la práctica. Los intervalos percentiles son menos erráticos, pero tienen coberturas de probabilidad satisfactorias.

Sea $\hat{\theta}^*(\alpha)$ el α -ésimo percentil de B replicaciones bootstrap: $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$. El intervalo percentil $1-2\alpha$ está dado por $(\hat{\theta}^*(\alpha), \hat{\theta}^{(1-\alpha)})$, los límites del intervalo están también dados por los percentiles de la distribución bootstrap pero no en la misma forma de (3.16). Los percentiles utilizados dependen de dos números \hat{a} y \hat{z}_0 , denominados la aceleración y la corrección del sesgo, respectivamente. El intervalo BCa de cobertura $1-2\alpha$, está dado por,

$$BCa : (\hat{\theta}^*(\alpha_1), \hat{\theta}^*(\alpha_2)) \quad (3.18)$$

donde,

$$\alpha_1 = \Phi \left[\hat{z}_0 + \frac{\hat{z}_0 + Z^{(\alpha)}}{1 + \hat{a}(\hat{z}_0 + Z^{(\alpha)})} \right] \quad (3.19)$$

$$\alpha_2 = \Phi \left[\hat{z}_0 + \frac{\hat{z}_0 + Z^{(1-\alpha)}}{1 + \hat{a}(\hat{z}_0 + Z^{(1-\alpha)})} \right] \quad (3.20)$$

Aquí, $\Phi(\cdot)$ representa la función de distribución normal estándar, $Z^{(\alpha)}$ y $Z^{(1-\alpha)}$ son los percentiles α y $1-\alpha$, respectivamente, de una distribución normal estándar. Notese que si \hat{a} y \hat{z}_0 entonces $\alpha_1 = \Phi(Z^{(\alpha)}) = \alpha$ y $\alpha_2 = \Phi(Z^{(1-\alpha)}) = 1 - \alpha$, de modo que el intervalo BCa se reduce al intervalo percentil.

¿Cómo se calculan \hat{a} y \hat{z}_0 ?

El valor *bias-correction* \hat{z}_0 se obtiene directamente de la proporción de replicaciones $\hat{\theta}^*$ que son menores que el estimador original $\hat{\theta}$, esto es,

$$\hat{z}_0 = \Phi^{-1} \left\{ \frac{\#\{\hat{\theta}^* < \hat{\theta}\}}{B} \right\} \quad (3.21)$$

Φ^{-1} indica la función inversa de una función de distribución normal estándar.

\hat{z}_0 mide aproximadamente el sesgo de la mediana de θ^* , que es la diferencia entre la mediana de $\hat{\theta}^*$ y $\hat{\theta}$.

Existen varias formas para calcular la aceleración \hat{a} . La explicación más fácil es dada en términos jackknife de un estadístico $\hat{\theta} = S(X)$. Sea $X_{(i)}$ la muestra original con el i-ésimo punto X_i borrado, sea $\hat{\theta}_{(i)} = S(X_{(i)})$ y se define

$$\hat{\theta}_{(.)} = \frac{\sum_{i=1}^n \hat{\theta}_{(i)}}{n} \quad (3.22)$$

Una expresión simple para la aceleración es

$$\hat{a} = \frac{\sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^3}{6 \left\{ \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2 \right\}^{3/2}} \quad (3.23)$$

La cantidad \hat{a} se denomina *aceleración* porque se refiere a la razón de cambio del error estándar de $\hat{\theta}$ con respecto al verdadero valor de θ . Valores no nulos de \hat{a} y \hat{z}_0 , cambian los percentiles utilizados para los límites *BCa*. Estos cambios corrigen ciertas deficiencias de los métodos *t-bootstrap*, estándar y percentil.

El intervalo *BCa* tiene la interesante propiedad de "transformación", es decir, los límites *BCa* se transforman correctamente si cambiamos el parámetro de interés de θ a alguna función de θ . Por ejemplo, el intervalo confidencial *BCa* para $\sqrt{V(A)} = \sqrt{\theta}$, se obtiene tomando las raíces cuadradas a los límites del intervalo *BCa* para θ . La principal desventaja del intervalo *BCa* es el número de replicaciones requeridas, al menos $B = 1000$ replicaciones son necesarias para reducir el error muestral de la simulación de Monte Carlo.

3.2.- METODO KERNEL

3.2.1.- Introducción

(Izenman, 1991) El campo de la estadística no paramétrica se ha enriquecido en años recientes con el desarrollo de nuevas herramientas para el análisis estadístico. Estas nuevas herramientas ofrecen alternativas a los tradicionales modelos paramétricos para explorar datos univariantes y multivariantes en grandes proporciones, sin necesidad de hacer grandes supuestos estructurales acerca de la distribución de los datos. Una de estas herramientas es la ESTIMACION NO PARAMETRICA DE DENSIDADES, la cual es un interesante tópico de investigación. //

(Silverman, 1986) ¿Qué es la estimación de una función de densidad de probabilidad (f.d.p.)? La f.d.p. es un concepto fundamental en Estadística. Considérese cualquier variable aleatoria X que tiene f.d.p. continua , donde

$$f(x) \geq 0, \int_{\mathbb{R}} f(x)dx = 1 \quad (3.24)$$

Si se especifica la forma funcional de f tenemos entonces una descripción de la distribución de X la cual nos permitirá, por ejemplo, obtener probabilidades asociadas con X a través de la relación

$$P(a < X < b) = \int_a^b f(x)dx \quad (3.25)$$

Supóngase ahora que no conocemos la f.d.p. y lo que se tiene es un conjunto de observaciones (muestra aleatoria) X_1, \dots, X_n de esta f.d.p. El

objetivo es estimar la f.d.p. desconocida , para lo cual se tienen dos enfoques:

(i) El Enfoque Paramétrico: este asume que los datos son seleccionados de una familia paramétrica de distribuciones, por ejemplo, la distribución normal de parámetros μ y σ^2 . Entonces la estimación de la densidad puede completarse, estimando los parámetros μ y σ^2 a partir de los datos observados, y sustituyendo luego estos valores en la f.d.p. normal.

(ii) El Enfoque No paramétrico: este es menos rígido en relación a los supuestos que se hacen acerca de la distribución de los datos observados, se asume que la distribución tiene f.d.p. f con al menos dos derivadas acotadas, y en este caso los datos observados nos pueden ayudar en la estimación de f en un mayor grado que en el caso de si fuese restringida a una familia paramétrica dada. En otras palabras, f , se supone que pertenece a una familia de densidades suficientemente grande, de modo tal que no puede representarse a través de un número finito de parámetros. //

(Izenman, 1991) Quizás el estimador no paramétrico más antiguo de una f.d.p. univariante fue el histograma. Progresos adicionales inicialmente - con los métodos kernel, series ortogonales, y el vecino más cercano - fueron inspirados por aplicación a la discriminación no paramétrica y desarrollos en estimación de densidad para series de tiempo estacionarias. Luego se desarrollaron métodos tales como verosimilitud penalizada, spline polinómicos, kernel variable, sierves, y projection pursuit. ¿ Por qué la estimación no paramétrica (y métodos relacionados) son populares hoy en día?. Se tiene una combinación de circunstancias: el crecimiento de la importancia de los computadores en la investigación estadística, la

disponibilidad de software estadístico de calidad y la ventajas de los gráficos de alta resolución.

Los investigadores han encontrado que la estimación no paramétrica de densidades es efectiva en la situaciones siguientes:

- (a) En análisis exploratorio, características descriptivas de la densidad estimada, tales como multimodalidad, comportamiento en las colas y asimetría son de especial interés; y el enfoque no paramétrico puede ser más flexible que los métodos tradicionales;
- (b) En análisis confirmatorio, los estimadores de densidad no paramétricos se utilizan en la toma de decisiones, tales como análisis discriminante no paramétrico y análisis de clasificación, pruebas para medianas, y pruebas para varianzas aleatorias; y
- (c) para propósitos de presentación, los detalles estadísticos de los datos con frecuencia pueden ser rápidamente explicados a individuos a través de simples gráficos de la curva de densidad estimada.

El éxito del desarrollo de la técnicas de estimación no paramétrica de densidad llevó, en su momento, a la formulación de la regresión no paramétrica, incluyendo el análisis no paramétrico de curvas de crecimiento, y patrones de reconocimiento estadístico no paramétricos.

3.2.2.- Estimadores de Densidad Confiables (Bona fide)

De los métodos disponibles para la estimación de densidades, algunos siempre nos suministran estimadores confiables, mientras que otros generalmente nos resultan en estimadores de densidad que contienen ordenadas negativas (especialmente en las colas) o tienen una integral infinita.

La negatividad puede ocurrir naturalmente, como resultado de que los datos se encuentran ubicados en ciertas regiones, o esto puede ser causado por la flexibilización de la restricción de no negatividad de (3.24) con el objeto de mejorar el porcentaje de convergencia de un estimador de f . Además, en la mayoría de los casos los investigadores para apresurar la convergencia de estimadores, seleccionan el flexibilizar la restricción (3.24) en lugar de la restricción de no negatividad.

Existen muchas otras formas de aliviar estos problemas: la densidad estimada puede truncarse a su parte positiva; alternativamente podría estimarse una versión transformada de f , digamos $\log f$ o $f^{\frac{1}{2}}$ y entonces efectuar la transformación necesaria para obtener un estimador no negativo de f .

3.2.3.- Propiedades Estadísticas de los Estimadores de Densidad

Como cualquier procedimiento estadístico, los estimadores no paramétricos de densidad son recomendados solamente si poseen propiedades deseables. Las propiedades para muestras finitas de los estimadores no paramétricos de densidad están disponibles para situaciones especiales (*Deheuvels 1977; Fryer 1976*); pero en general, la investigación se ha centrado en el desarrollo de propiedades para muestras grandes.

Insesgabilidad: Un estimador \hat{f} de una f.d.p. f es insesgado si $\forall x \in \mathbb{R}, E_f[\hat{f}(x)] = f(x)$. Aunque existen estimadores insesgados de densidades paramétricas tales como la normal, poisson, exponencial, y geométrica. No existe un estimador confiable para todas las densidades continuas que sea insesgado.

Por tanto, la atención se ha enfocado sobre sucesiones $\{\hat{f}(x)\}$ de estimadores de densidades no paramétricos que son asintóticamente insesgados para f , esto es, $\forall x \in \mathbb{R}, E_f[\hat{f}(x)] \rightarrow f(x)$ cuando $n \rightarrow \infty$.

Consistencia: Una propiedad más importante es la consistencia, f es un estimador (débilmente) consistente puntual para una f.d.p. f univariante si $\hat{f}(x) \rightarrow f(x)$ en probabilidad $\forall x \in \mathbb{R}$, y es fuertemente consistente puntual para si converge casi seguramente.

El enfoque L_2 . Si asumimos que f es cuadrado integrable, entonces el desarrollo de \hat{f} en $x \in \mathbb{R}$ se mide por el **Error Cuadrático Medio**.

$$MSE(x) = E_f[\hat{f}(x) - f(x)]^2 = V[\hat{f}(x)] + \{Sesgo[\hat{f}(x)]\}^2 \quad (3.26)$$

donde,

$$\begin{aligned} V[\hat{f}(x)] &= E_f\{\hat{f}(x) - E_f[\hat{f}(x)]\}^2 \\ Sesgo[\hat{f}(x)] &= E_f[\hat{f}(x)] - f(x) \end{aligned}$$

Si $MSE(x) \rightarrow 0$ cuando $n \rightarrow \infty$ entonces se dice que \hat{f} es un estimador puntual consistente de en media cuadrática.

Un criterio más importante relaciona las curvas \hat{f} y f . Una de estas medidas de ajuste se encuentra integrando (3.26) sobre todos los valores de x , esto nos resulta en el error cuadrático medio integrado (*MISE*),

$$MISE = \int_{\mathbb{R}} E_f[\hat{f}(x) - f(x)]^2 dx \quad (3.27)$$

Otra medida comúnmente empleada es el Error cuadrado integrado o norma L_2

$$ISE = \int_{\mathbb{R}} [\hat{f}(x) - f(x)]^2 dx \quad (3.28)$$

Tomando esperanza sobre en (3.28) nos resulta en el *MISE*. Con frecuencia el *ISE* se prefiere como criterio en lugar de su valor esperado, *MISE*, ya que el *ISE* determina qué tan próximo está \hat{f} de f en un conjunto de datos,

mientras que el *MISE* esta referido con el promedio de todos los conjuntos de datos posibles.

Bajo condiciones poco rigurosas, se ha demostrado que el *ISE* es una razonable aproximación aleatoria del *MISE* (*Marron and Hardle, 1986*), mientras que en ciertas situaciones, el *MISE* puede ser un mejor criterio que el *ISE* (*Hall and Marron 1988*), *Farrell* (1972) demostró que para los estimadores de densidad confiables la mejor proporción de convergencia asintótica de *MISE* es $O(n^{-\frac{4}{5}})$, y *Boyd and Steele* (1978) demostraron que $O(n^{-1})$, si f es una densidad normal.

El enfoque L_1 . Un problema con el enfoque L_2 para estimación de densidades no paramétrica es que el comportamiento en el extremo de una densidad viene a ser menos importante, posiblemente resultando en peculiaridades en los extremos de la densidad estimada. Un enfoque alternativo para la teoría no paramétrica de estimación de densidades es L_1 . Específicamente *Devroye and Gyorfi* (1985) expresaron que L_1 es "el espacio natural para las densidades", y demostraron que el error absoluto integrado (también conocido como la variación total o norma L_1),

$$LAE = \int_{\Omega} |\hat{f}(x) - f(x)| dx \quad (3.29)$$

esta siempre definido como una norma sobre ese espacio, es invariante bajo transformaciones monótonas, y $0 \leq LAE \leq 2$. Si $LAE \rightarrow 0$ en probabilidad cuando $n \rightarrow \infty$, entonces se dice que \hat{f} es un estimador consistente de f ; la consistencia fuerte de \hat{f} ocurre cuando converge casi seguramente. Al tomar el valor esperado de (3.29) sobre todas las densidades nos resulta en el error medio absoluto integrado, $MIAE = E_f[LAE]$.

Una cosa si es clara la labor técnica para obtener L_1 resulta sustancialmente más difícil que la necesaria para obtener su análogo resultado en L_2 . //

3.2.4.- Estimadores Kernels

3.2.4.1.- Conceptos Informales de Estimación Kernel

(Hand, 1982) Para introducir los conceptos fundamentales de una manera intuitiva, comenzaremos ilustrando formas diferentes de búsqueda de tales estimadores.

Sea una muestra X_1, \dots, X_n , nuestro objetivo es encontrar un estimador $\hat{f}(x)$, de la f.d.p. condicional f en un punto x basado en X_1, \dots, X_n . //

(Silverman, 1986) La forma más sencilla es a través del histograma. Dado un origen x_0 y un intervalo de tamaño h , se definen los intervalos del histograma de la forma siguiente $[x_0 + mh, x_0(m+1)h]$ para un entero m positivo o negativo. Por tanto el estimador de densidad esta dado por

$$\hat{f}(x) = (nh)^{-1} \{ \# \text{ de } X_i \text{ en un intervalo dado} \} \quad (3.30) //$$

(Hand, 1982) El histograma presenta una serie de desventajas, entre las cuales están la naturaleza fija de los intervalos - la cual puede remediararse mediante intervalos de amplitud diferente - las discontinuidades en los límites de los intervalos, y el hecho de que sea cero en cierto rango. La primera de estas propiedades significa que aquellos puntos algo distanciados pero dentro del mismo intervalo tienen el mismo estimador de densidad mientras que puntos cercano a otros pero en diferentes lados de un límite de intervalo pueden tener muy diferentes estimadores.

En 1956 Murray Rosenblatt utilizó ideas análogas a aquellas de los promedios móviles en análisis de series de tiempo para aliviar estos problemas, en lugar de tener una estructura de intervalos fija,

independientemente de la posición de x , el punto en el cual deseamos estimar la densidad, el sugirió centrar el intervalo en x . En un histograma la probabilidad de que un punto caiga en el intervalo centrado en x es estimada por la proporción de puntos que caen en ese intervalo. La densidad de probabilidad en x es en este caso la probabilidad total dividida por el volumen del intervalo (en nuestro caso el volumen es la longitud ya que corrientemente consideraremos sólo una variable). De este modo si el intervalo es de longitud $2h$:

$$\hat{f}(x) = n^{-1} \sum_{i=1}^n \frac{\mathbf{K}_1(x-x_i)}{2h} \quad (3.31)$$

donde, $\mathbf{K}_1(u) = \mathbf{I}_{[-h,h]}(x)$.

Entonces, es un estimador más representativo de la muestra.

Otra forma de ver a (3.31) nos lleva a la idea de considerar (3.31) como un promedio de n valores, uno por cada uno de la muestra. Los puntos cercanos a x (dentro de la distancia h) contribuyen un valor $\frac{1}{2h}$ mientras que los puntos fuera de la distancia h entonces contribuyen con un valor cero. Claramente el salto de $\frac{1}{2h}$ a 0 es el que lleva a discontinuidades, así, podríamos sustituir $\frac{\mathbf{K}_1}{2h}$ por otra función $\mathbf{K}_2(x-x_i)$ la cual decrece gradualmente con el incremento de $|x-x_i|$. Nuestro estimador sería entonces

$$\hat{f}(x) = n^{-1} \sum_{i=1}^n \mathbf{K}_2(x-x_i) \quad (3.32)$$

donde $\mathbf{K}_2(u)$ decrece monótonamente y sin saltos como $|u|$ se incrementa (más generalmente, $\mathbf{K}_2(x)$ puede cambiar gradualmente y no es monótona decreciente). Ahora en regiones con muchos x'_i s habrá mayor

contribución, mientras que en regiones con muy pocas x_i 's entonces habrá muy poca contribución.

Note que si \mathbf{K}_2 tiene como rango la recta real, entonces la tercera desventaja del histograma - la ausencia de colas - es suprimida (aunque no completamente, ya que eso dependerá del \mathbf{K}_2 en particular seleccionado)

Retornando por el momento al kernel Rosenblatt de (3.31) este puede reescribirse como

$$\hat{\mathbf{f}}(x) = (2nh)^{-1} \{F_n(x+h) - F_n(x-h)\} \quad (3.33)$$

donde $F_n(x)$ es la función de distribución muestral,

$$\begin{aligned} f(x) &= \lim_{h \rightarrow 0} (2nh)^{-1} P(x-h < X < x+h) \\ &= \lim_{h \rightarrow 0} (2h)^{-1} \{F_X(x+h) - F_X(x-h)\} \end{aligned} \quad (3.34)$$

ya que la f.d.p. es la derivada de la función de distribución. Aparentemente para n fijo un valor grande de h resulta en un estimador suavizado, mientras que un valor pequeño de h resulta en un estimador altamente irregular. Si h es suficientemente grande el conjunto X_1, \dots, X_n estará dentro del intervalo y resulta entonces un estimador uniforme. En el límite cuando $h \rightarrow 0$, $\hat{\mathbf{f}}$ corresponderá a la derivada de la F.D. muestral, entonces $\hat{\mathbf{f}}$ será una serie de picos de probabilidad cada uno situado en un x_i . La forma general de (3.32) puede reescribirse como

$$\hat{\mathbf{f}}(x) = \int \mathbf{K}(x-y)dF_n(y) \quad (3.35)$$

lo cual nos indica que $\hat{\mathbf{f}}$ es una convolución suavizada de la F.D muestral.

Aquí aparece otra desventaja asociada con el histograma y que no mencionamos anteriormente, esta es el problema de selección del tamaño de los intervalos. Desafortunadamente la adopción del método kernel no resuelve esta dificultad. Es evidente que el tamaño de los intervalos debe seleccionarse de modo tal que estas irregularidades puedan modelarse. Esto

también depende del tamaño de la muestra, n . Vamos a denominarlo *parámetro de suavización - tamaño de ventana, amplitud de banda* - denotado por h , para variables continuas. Como es necesario especificar la dependencia del estimador sobre h , adoptaremos la forma general

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) \quad (3.36)$$

En esta circunstancia, podemos utilizar solamente información local acerca del valor de la densidad en algún punto dado x . Es decir, el valor de la densidad en un punto x debe calcularse de los valores X_1, \dots, X_n que caen en una vecindad de x , y para asegurar la consistencia, la vecindad debe contraerse cuando el tamaño de la muestra se incrementa. En el caso del estimador kernel, el radio de efectividad de la vecindad es aproximadamente igual al "tamaño de banda" o "parámetro de suavización" del estimador.

Como nota final debemos comentar que para nuestros propósitos, el método kernel elimina otra desventaja del histograma básico. Esta no mencionada antes es el incremento exponencial en el número de intervalos, como el número de variables, se incrementa. Este problema se conoce como "el curso de dimensionalidad". El estimador kernel es siempre un promedio de n contribuciones, no importando cuántas variables están involucradas. //

(Zelenman, 1991) Fue Cacoullos, 1966, el primero en denominar a K en (3.36) una función kernel, previamente a K se le denominaba función de ponderación. En el caso univariante, la transformada rápida de Fourier se recomienda para efectos de cálculo ya que se demuestra que \hat{f}_h conserva las propiedades inherentes al kernel K .