ANALISIS DE SISTEMAS

¿ ES EL METODO KERNEL* SUPERIOR A LOS HISTOGRAMAS?

Isaac Dyner. – Profesor Asociado

UNI **Universidad Nacional de Colombia – Seccional Medellín**

FACULTAD DE MINAS

MEDELLIN, 1981

1. INTRODUCCION

En muchos problemas de casi todos los campos de las Ciencias Aplicadas, aparece la necesidad de estimar funciones de densidad f(.) de una variable aleatoria X. Dentro de los métodos de estimación de densidades, el Histograma es el más antiguo y, posiblemente, el más utilizado en la actualidad. Esto último se debe principalmente a su simpleza y bondad.

La primera referencia de uso del Histograma se encuentra en el trabajo de John Graunt (ver [6]) presentado ante la Sociedad Real de Inglaterra en 1661. Desde ese entonces poco se avanzó, hasta la aparición en 1895 de la Familia de Distribuciones de Karl Pearson [3]. A pesar de que en el presente siglo han aparecido otros métodos como el de Máxima Verosimilitud, Familias de Johnson y Series, es sólo a partir del surgimiento del método Kernel en 1956 cuando se genera un nuevo progreso en la estimación de funciones de densidad.

El autor intenta hacer en este trabajo una introducción que motive el uso del método de los Kerneles el cual, con el avance de los computadores, puede ser de gran utilidad para los investigadores en las Ciencias Aplicadas.

---

* En literatura matemática se ha traducido el término Kernel por el de núcleo.

En lo subsiguiente sólo se considera la estimación de funciones de densidad absolutamente continuas en un intervalo finito de la forma [a, b].

2. HISTOGRAMAS

Para la elaboración de un Histograma con base en una muestra $X_1, \ldots Xn$ sobre un intervalo (a, b) "normalmente" se procede de la siguiente forma:

Primero se hace una partición de (a, b) en m subintervalos iguales, de longitud (b−a)/m, y finalmente se determina la proporción de los Xi's en cada uno de los subintervalos. En una forma más general, un Histograma está definido por una partición $a = t_0 < t_1 \ldots < t_m = b$ y una función fH (.) tal que

$$(1) \quad f_H(t) = \begin{cases} C_i & t_i \leq t < t_{i+1} \quad i = 0,1,\ldots,m-1 \\ C_{m-1} & t = b \\ 0 & t \notin [a, b] \end{cases}$$

Con $f_H(t) \geq 0$, $\quad -\infty < t < \infty$

y $$\int_a^b f_H(t)\, dt = 1$$

Se puede demostrar (ver [5]) que los estimadores de máxima verosimilitud de los parámetros $C_0, \ldots, C_n$ de (1) están dados por

$$\hat{C}_i = \frac{q_i}{n(t_{i+1} - t_i)} \quad , i = 0, 1, \ldots, m-1$$

Donde n es el tamaño de la muestra y q_i representa el número de observaciones en el i-ésimo intervalo.

También se puede demostrar (ver [5]) que bajo ciertas condiciones generales, f_H (.) es un stimador consistente de f (.). Es decir

$$(2) \quad E\,[\,(f_H(x) - f(x))^2\,] \longrightarrow 0 \qquad \text{cuando} \quad n \longrightarrow \infty$$

Es importante anotar que para una partición óptima del intervalo [a, b] la velocidad de convergencia de (2) es del orden de $n^{-2/3}$.

3. ESTIMADORES KERNEL

Como pudimos observar en la sección anterior, los Histogramas no son "malos" estimadores de las funciones de densidad. Sinembargo, $f_H(.)$ es discontinuo, completamente Ad-hoc, difícil de actualizar y requiere de muchos parámetros. Una generalización del Histograma se debe a Rosenblatt [4], quien propuso el estimador

$$(3) \qquad F_n(x) = \frac{F_n(x + h_n) - F_n(x - h_n)}{2h_n}$$

Donde $h_n \in R \quad , \quad \forall n \quad y$

$$(4) \qquad F_n(x) = \frac{\text{No. de puntos muestrales} \leq x}{n}$$

Para este estimador, la velocidad de convergencia en (2), cuando se sustituye $f_H(\cdot)$ $g.:f$ /
ye $f_H(.)$ por $f_n(.)$ es del orden de $n^{-4/5}$ (más veloz que la del Histograma),

$$h_n = \left[\frac{9 \quad f(x)}{2\,(f''(x)^{-})^2}\right]^{1/5} n^{-1/5}$$

En general no se puede encontrar h_n ya que está expresado en función de f(.) que es precisamente la densidad que se pretende estimar. Sinembargo, en la práctica, se hacen diferentes ensayos hasta obtener un Histograma que se ajuste a la muestra.

Con el estimador (3) se sigue teniendo muchos de los obstáculos que se tenían nían con los Histogramas clasicos, como la discontinuidad y la forma arbitraria de determinarlos.

E. Parzen [2] propone una generalización del estimador (3) de Rosenblatt. Este está dado por

$$(5) \qquad \hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - Xi}{h_n}\right)$$

La función K(.) debe cumplir las siguientes propiedades:

$$K(x) \geq 0, \quad -\infty < x < \infty \qquad \text{y} \qquad \int_{-\infty}^{\infty} K(x)\,dx = 1$$

Además

$$\text{Sup } K(x) < \infty$$

$$-\infty < x < \infty$$

y $$\lim_{x \to \infty} xK(x) = 0$$

A K(.) se le denomina **Kernel**, y por esto a $f_n(.)$ se le llama **estimador Kernel.**

Se puede observar que este estimador cumple con las propiedades de una función de densidad, es decir:

(a) $\hat{f}_n(x) \geq 0 \quad \forall x,$

y (b) $$\int_{-\infty}^{\infty} \hat{f}_n(x)\,dx = \frac{1}{nh_n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{x - xi}{h_n}\right)dx$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K(y)\,dy = 1$$

Además los estimadores $f_n(.)$ son asintóticamente insesgados y consistentes en el sentido (2) (con una velocidad de convergencia del orden $n^{-2r/(2r-1)}$, para escogencias

para escogencias óptimas de h_n y de un entero positivo r que depende de ciertas propiedades de K(.), [5]).

Algunos de los Kerneles más conocidos son:

(6) $$K(x) = \begin{cases} 1/2 & |x| \leq 1 \\ 0 & |x| > 1 \end{cases}$$

$$(7)\quad K(x) = \begin{cases} 1 - x & |x| \leq 1 \\ 0 & |x| > 1 \end{cases} \qquad y$$

$$(8)\quad K(x) = \frac{1}{(2\pi)^{1/2}} \; Exp\left(-\frac{1}{2} x^2 \right) \qquad -\infty < x < \infty$$

Nótese que el Kernel (6) genera el estimador (3) definido por Rosenblatt ya que

$$\hat{f}_n(x) = \frac{1}{n h_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right)$$

$$= \frac{1}{2n h_n} \; \# \left\{ X_i / x - h_n < X_i < x + h_n \right\}$$

$$= \frac{\# \text{ de puntos muestrales en } (x - h_n, x + h_n)}{2n h_n}$$

$$= \frac{F_n(x + h_n) - F_n(x - h_n)}{2 h_n} = f_n(x)$$

La literatura escrita acerca del estimador Kernel es bastante extensa y aquí sólo hemos esbozado algunas de las propiedades más elementales de este estimador. Para un estudio más profundo se recomienda el libro de Tapia y Thompson [5]; para un análisis de lo publicado en este campo el artículo de Fryer [1] puede ser orientador.

4. EJEMPLO

Con base en una muestra del peso de cuarenta recién nacidos en un hospital de Medellín (ver Tabla 1), se ilustra por medio de las Figuras 1, 2, 3 y 4 el gráfico del estimador Kernel (8) para diferentes valores de su parámetro h_n. Se puede observar como en la medida en que h_n crece el estimador es más suave. En la Figura 3 se pueden observar conjuntamente un Histograma y un estimador Kernel.

TABLA 1

2300	2350	2450	2500	2500	2500	2500	2600
2610	2650	2650	2700	2840	2900	2900	2900
2900	2900	2950	2950	3000	3000	3040	3080
3100	3100	3100	3100	3150	3150	3300	3300
3300	3350	3350	3400	3450	3500	3550	3800

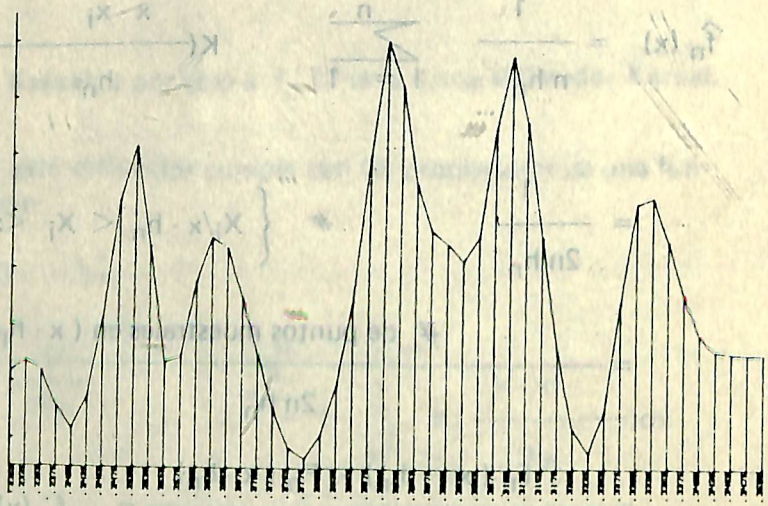


Figura 1

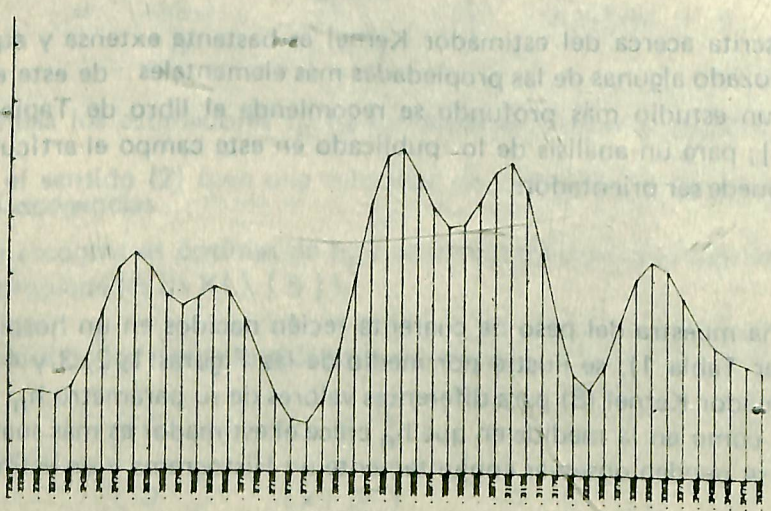
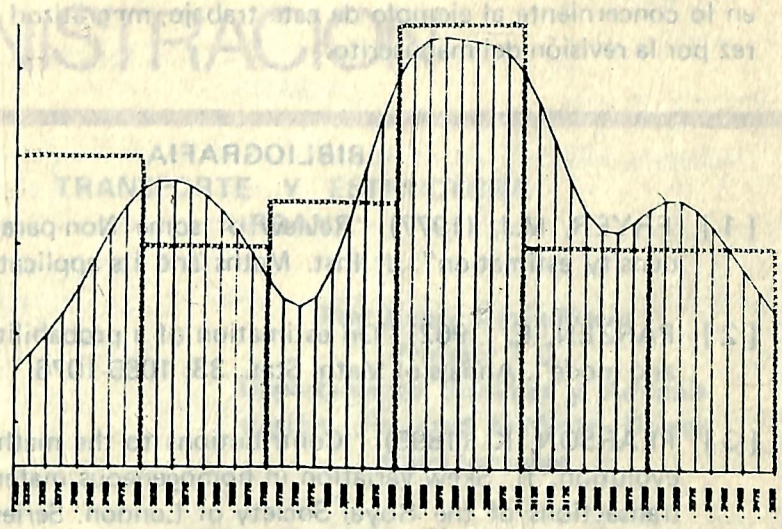


Figura 2

Figura 3



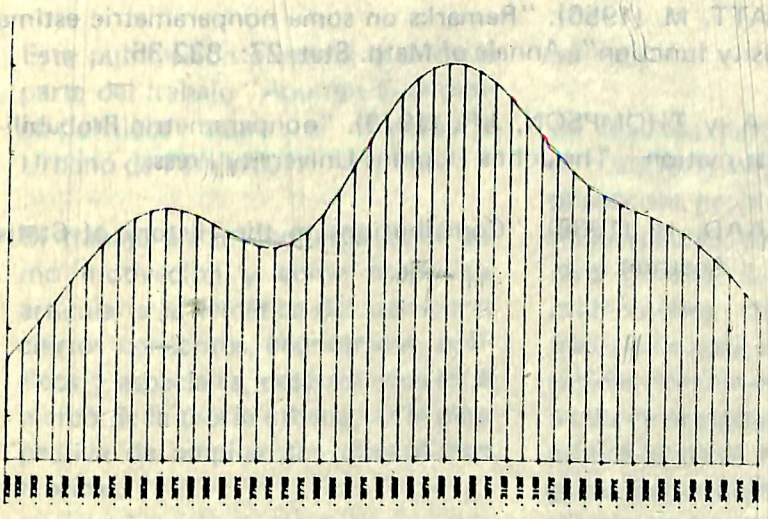


Figura 4

CONCLUSIONES

Se ha presentado un método relativamente nuevo (y no conocido en nuestro medio) de estimación de funciones de densidad. El estimativo $f_n(.)$ expuesto aquí es continuo, diferenciable en todas partes, fácil de actualizar y con un solo parámetro por estimar.

AGRADECIMIENTO

Quiero agradecer a Alvaro López y a Luis Carlos López por la ayuda prestada

en lo concerniente al ejemplo de este trabajo, mi gratitud al profesor Luis Pérez por la revisión del manuscrito.

BIBLIOGRAFIA

[1] FRYER, M.J. (1977) "Review of some Non-parametric methods of density estimation". J. Inst. Maths and its applications. 20: 335-354.

[2] PARZEN, E. (1962) "On estimation of a probability density function and mode". Annals of Math. Stat. 33: 1065-1076.

[3] PEARSON, K. (1895). "Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material". Philosophical transactions of the Royal Society of London. Series A 186: 343-414.

[4] ROSENBLATT, M. (1956). "Remarks on some nonparametric estimates of a density function". Annals of Math. Stat. 27: 832-35.

[5] TAPIA, R. A. y THOMPSON, J.R. (1978). "nonparametric Probability density estimation". The Johns Hopkins University Press.

[6] WESTERGAAD, H. (1968). "Contributions to the History of Statistics". N. Y. : Agathon.