# Estadística III

## Estimación de la Densidad

Alejandro López Hernández

FES Acatlán - UNAM

May 16, 2020

# Índice

1. Introducción
2. Histogramas
3. Estimación con Kernel

El objetivo de la estimación no paramétrica de la función de densidad es hacer una cantidad mínima de supuestos. Supongamos que tenemos una muestra aleatoria $X_1, ..., X_n \sim F$, entonces nuestro objetivo es estimar $f = F'$. Nuestra estimación la denotaremos como $\hat{f}_n$, y aparte de que dependa de $n$ también dependerá de un parámetro llamado *ancho de banda*.

La forma más simple de estimar la densidad son los histogramas. La idea de los histogramas es aproximar $f$ mediante una función escalonada, el tamaño de los escalones se determinara por la cantidad de casos en cada uno de los intervalos que se haya definido para particionar el rango de las variables aleatorias. De forma mas precisa, supongamos que tenemos una variable aleatoria cuyo rango es el intervalo $[0, 1]$, entonces lo dividimos en $m$ particiones iguales

$$B_1 = \left[0, \frac{1}{m}\right), \quad B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), ..., \quad B_m = \left[\frac{m-1}{m}, 1\right)$$

De esta manera definimos el ancho de banda como $h = \frac{1}{m}$, los conjuntos $B_j$ son llamados *bins*.

Sea $Y_j$ la cantidad de observaciones en $B_j$, y sea $p_j = \int_{B_j} f(u)du$ cantidad que podemos estimar como $\hat{p}_j = Y_j/n$, entonces nuestro histograma se define como:

$$\hat{f}_n(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} 1_{B_j}$$

La motivación es la siguiente:

$$\mathbb{E}(\hat{f}_n(x)) = \frac{\mathbb{E}(\hat{p}_j)}{h} = \frac{\hat{p}_j}{h} = \frac{\int_{B_j} f(u)du}{h} \approx \frac{f(x)h}{h} = f(x)$$

**Teorema 1**

Supongamos que $f'$ es absolutamente continua y que $\int (f(u))^2 du < \infty$. Entonces

$$R(\hat{f}_n(x), f) = \frac{h^2}{12} \int (f(u))^2 du + \frac{1}{nh} + o(h^2) + o\left(\frac{1}{n}\right)$$

El valor optimo para el ancho de banda es

$$h^* = \frac{1}{n^{1/3}} \left( \frac{6}{\int (f(u))^2 du} \right)^{1/3}$$

, si escogemos ese ancho de banda entonces

$$R(\hat{f}_n(x), f) \sim \frac{(3/4)^{2/3} (\int (f(u))^2 du)^{1/3}}{n^{2/3}}$$

La métrica para medir la calidad de nuestra estimación $\hat{f}_n$, sera con la error que cometemos al hacer la estimación con un ancho de banda de $h$,

$$L(h) = \int (\hat{f}_n(u) - f(u))^2 du$$

Derivado de esta cantidad

Definición

Se define el estimar de **riesgo de validación cruzada** como

$$\hat{J}(h) = \int (\hat{f}_n(x))^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)$$

donde $\hat{f}_{-i}$ es la estimación de la densidad sin considerar la i-ésima observación.

Para la estimación de histogramas tenemos que

Teorema 2

$$\hat{J}(h) = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)} \sum_{j=1}^{m} \hat{p_j}^2$$

Una de las desventajas de la estimación mediante histogramas es que no es una función suave y que la velocidad de convergencia. La estimación con Kernel soluciona estos 2 problemas y resulta ser la estimación que mas rápido converge a la densidad real. Esta estimación propone una suma de funciones **Kernel**, una Kernel es una función suave mayor a 0 tal que

$$\int K(x)dx = 1, \quad \int xK(x)dx = 1, \quad \int x^2K(x)dx > 0$$

Algunos ejemplos de funciones kernel son:

- $K(x) = \frac{1}{2} 1_{|x| \leq 1}$
- $K(x) = \frac{1}{\sqrt{2\pi}} e^{x^2/2}$
- $K(x) = \frac{3}{4}(1 - x^2) 1_{|x| \leq 1}$
- $K(x) = \frac{70}{81}(1 - |x|^3)^3 1_{|x| \leq 1}$

### Definición

Dado un kernel $K$ y un valor positivo $h$, a estimación por kernel de la densidad esta definida como

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

Tenemos los siguientes 2 resultados relevantes

### Teorema 3

Supongamos que $f$ es continua en $x$ y que $h_n \to 0$ y $nh_n \to \infty$ cuando $n \to \infty$. Entonces $\hat{f}_n(x) \to f$ en probabilidad.

### Teorema 4

Para cualquier $h > 0$

$$\hat{J}(h) = \frac{1}{hn^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K^* \left( \frac{X_i - X_j}{h} \right) + \frac{2}{nh} K(0) + O\left( \frac{1}{n^2} \right)$$

Donde $K^*(x) = K^{(2)}(x) - 2K(x)$ y $K^{(2)}(z) = \int K(z-y)K(y)dy$