Probabilidad y Estadística

Clase 4

Estadística

¿Qué es la estadística?

Ahora lo que tenemos son observaciones (realizaciones) de las variables aleatorias.

El objetivo es poder hacer algún tipo de inferencia a partir de los valores observados.

Muestra aleatoria

Supongamos que tenemos un experimento aleatorio relacionado con una v.a. X. La v.a. X representa un observable del experimento aleatorio.

Los valores de X son la población de estudio, y el objetivo es saber como se comporta esa población.

Una muestra aleatoria de tamaño n, es una sucesión de n v.a **independientes** $\underline{X}_n = (X_1, \ldots, X_n)$, tal que $X_i \sim X$

Estimador

Def: Un estimador para una cierta magnitud θ (desconocida) de la distribución es una función $\delta(\underline{X})$ de la muestra aleatoria, que un valor aproximado de

Error cuadrático medio

Def: El error cuadrático medio (ECM) como $\mathbb{E}[(\delta(\underline{X}_n) - \theta)^2]$

Def: Un estimador $\delta^*(\underline{X})$es óptimo si
$ECM(\delta^*(\underline{X})) \leq ECM(\delta(\underline{X}))$ para todo $\delta(\underline{X})$.

Bondades de los estimadores

Def: Diremos que $\delta(\underline{X})$ es un estimador insesgado para θ si $B = \mathbb{E}[\delta(\underline{X}) - \theta] = 0 \quad \forall\theta$. En caso contrario diremos que es sesgado. A B se lo conoce como $\delta(\underline{X})$ sesgo.

Def: Diremos que $\delta(\underline{X})$ es un estimador asintóticamente insesgado para θ si $\lim_{n\to\infty} \mathbb{E}[\delta(\underline{X}_n) - \theta] = 0 \quad \forall\theta$

Obs: El ECM se puede descomponer como:

$$ECM = \underbrace{var(\delta(\underline{X}_n))}_{varianza} + \underbrace{B(\delta(\underline{X}_n))^2}_{sesgo}$$

Bondades de los estimadores

Def: Dada una sucesión de estimadores $\delta(\underline{X}_n)$ de θ, diremos que $T = \delta(\underline{X})$ es (débilmente) consistente si

$$\forall\, \varepsilon > 0, \mathbb{P}(|T - \theta| > \varepsilon) \to 0$$

Teorema: Si $var(\delta(\underline{X})) \to 0$ y $\mathbb{E}[\delta(\underline{X})] \to \theta$, entonces $\delta(\underline{X})$ es consistente.

Def: Un estimador es consistente en media cuadrática si

$$\lim_{n\to\infty} ECM(\delta(\underline{X}_n)) = 0, \forall\theta$$

Estimadores de mínimos cuadrados

Estimador de mínimos cuadrados

Es común querer estimar el valor de una v.a. X a partir de una medición Y. Ejemplo: Y es una versión ruidosa de X.

Buscamos un estimador $\hat{X}$ de X tal que tenga mínimo error cuadrático medio

$$ECM = \mathbb{E}[(X - \hat{X})^2]$$

Observar que se corresponde con la distnacia asociada al p.i. canónico para v.a.

Estimador de mínimos cuadrados

En otras palabras, queremos $\hat{X} = g^*(Y)$ tal que

$$\mathbb{E}[(X - \hat{X})^2] \leq \mathbb{E}[(X - g(Y))^2] \quad \forall\, g(Y) \text{ (medible).}$$

¿Quién era $\hat{X}$? $\hat{X} = \mathbb{E}[X|Y]$

Idea de demostración: [Ejercicio]

1. Probar que el mejor estimador constante es $\mathbb{E}[X]$
2. Probar que el mejor estimador condicional es $\mathbb{E}[X|Y = y]$.
3. Dejar que Y tome todos los valores posibles (i.e. reemplazo y por Y), recupero la esperanza condicional.

Mínimos cuadrados: caso lineal

A veces obtener $\mathbb{E}[X|Y]$ puede ser muy complicado, entonces nos restringimos a los estimadores lineales.

Buscamos a, b tq $\mathbb{E}[(X - (aY + b))^2]$ sea mínima.

Resulta que $a = \frac{cov(X,Y)}{var(Y)}$ y $b = \frac{cov(X,Y)}{var(Y)}\mathbb{E}[Y] + \mathbb{E}[X]$

La demostración la hicimos en Análisis Matemático

Regresión lineal

Tengo las observaciones $(x_1, \ldots, x_n)$ e $(y_1, \ldots, y_n)$, observaciones de dos v.a. X e Y, y queremos hallar la mejor relación lineal $Y = aX + b$.

Definiendo $A = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$, tenemos que

$[a, b]^T = (A^T A)^{-1} A^T y$ Nuevamente, la demostración la vimos en AM.

Estimadores puntuales

Estadístico

Def: Dada una muestra aleatoria $\underline{X}_n$, un estadístico es cualquier función $T_n = T(\underline{X}_n)$

Def: Sea una muestra aleatoria $\underline{X}_n$, cuya distribución es $F_\theta(\underline{x}), \theta \in \Theta$, se dice que $T = r(\underline{X}_n)$ es un estadístico suficiente para θ si $F_{\underline{X}|T=t}(\underline{x})$ no depende de θ.

Teorema de factorización: Diremos que $T = r(\underline{X}_n)$ es un est. suficiente para θ sii existen funciones h y g tales que:

$f_\theta(\underline{x}) = g(r(\underline{x}), \theta)h(\underline{x})$

Estimadores puntuales

Def: Dada una muestra aleatoria $\underline{X}_n$, un estimador $\hat{\theta}$ de un parámetro θ es una función de la muestra aleatoria que provee un valor aproximado del parámetro o característica desconocido.

Método de Máxima Verosimilitud

Def: Diremos que $\hat{\theta}(\underline{X})$ es un Estimador de Máxima Verosimilitud (MLE) si se cumple que:

$$f(\underline{X}, \hat{\theta}) = \max_{\theta} f_{\theta}(\underline{X})$$

La idea es que si observé una determinada muestra, entonces esta debería tener alta probabilidad de ocurrir, por lo tanto busco el θ que maximiza esa probabilidad de ocurrencia

Método de Máxima Verosimilitud

Def: Definimos la función de verosimilitud como $L(\theta) = f(\underline{x}, \theta)$ (vista como función de θ) luego,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

Si el soporte de X no depende θ, Θ es un conjunto abierto y $f_\theta(x)$ es derivable respecto de θ, entonces para hallar el EMV puedo hallar θ tal que

$$\frac{\partial\, ln(L(\theta))}{\partial\, \theta} = 0$$

Principio de invariancia

Supongamos que ahora queremos estimar por máxima verosimilitud a $\lambda = q(\theta)$.

Teorema: Si $\hat{\theta}$ es MLE de θ, entonces $\hat{\lambda} = q(\hat{\theta})$.

¿Por qué es útil? Por ejemplo, podría querer estimar una probabilidad de la v.a. X, que en general no puedo porque desconozco el parámetro de la distribución.

Estimación no paramétrica

Función de distribución empírica

Def: Sea $\underline{X}_n$ una m.a. tal que $X_i \overset{i.i.d}{\sim} F$, donde F es una función de distribución. La es función de distribución empírica (ECDF) es una función $\widehat{F}_n$ que pone masa $1/n$ en cada observación X_i.

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I\{X_i \leq x)\}}{n}$$

Propiedades de la ECDF

$$
\begin{aligned}
\mathbb{E}\left(\widehat{F}_n(x)\right) &= F(x), \\
\mathbb{V}\left(\widehat{F}_n(x)\right) &= \frac{F(x)(1-F(x))}{n}, \\
\text{MSE} &= \frac{F(x)(1-F(x))}{n} \to 0, \\
\widehat{F}_n(x) &\xrightarrow{\text{P}} F(x).
\end{aligned}
$$

All of Statistics, Wasserman

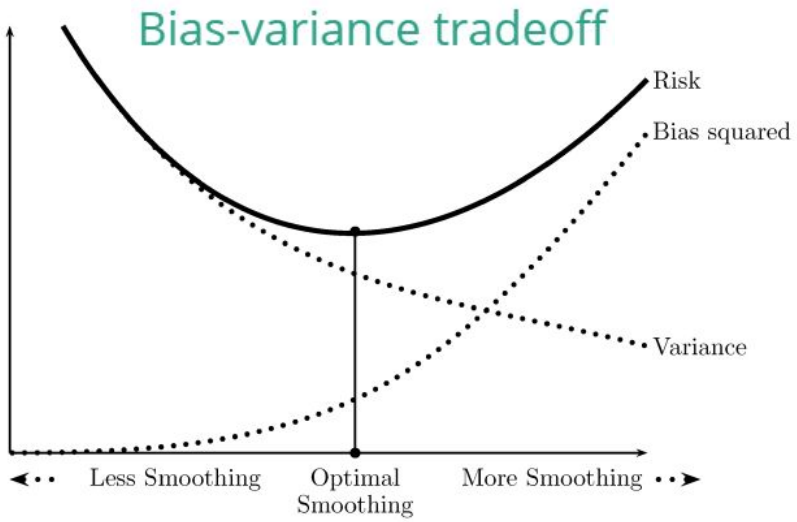
Estimación de densidades (*smoothing*)

Si deseamos estimar una función de densidad $f(x)$ o una función de regresión $\phi(x) = \mathbb{E}[X|Y = y]$, se deben hacer algunas suposiciones de suavidad.

Sea $\hat{g}_n$ un estimador de g.

Definimos el riesgo (error cuadrático medio integrado (MISE)) como

$$R(g, \hat{g}_n) = \mathbb{E}\left[\int g(u) - \hat{g}_n(u)\, du\right]$$



All of Statistics, Wasserman

Histogramas

1. Se selecciona un origen x_0 y se divide la recta real en intervalos de longitud h

$$B_j = [x_0 + (j-1)h, x_0 + jh], j \in \mathbb{N}$$

2. Se cuenta cuantas observaciones caen en cada intervalo armando una tabla de frecuencias. Denotamos a la cantidad de observaciones que caen en el intervalo j como n_j
3. Para cada intervalo, se divide la frecuencia absoluta por la cantidad total de la muestra n (para convertirlas en frecuencias relativas, análogo a como se hace con las probabilidades) y por la longitud h (para asegurarse que el area debajo del histograma sea igual a 1):

   Formalmente, el histograma está dado por:

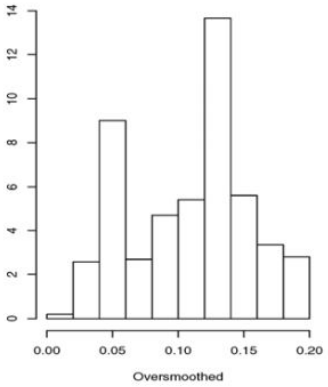
$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j} \mathbf{1}(x_i \in B_j)\mathbf{1}(x \in B_j)$$

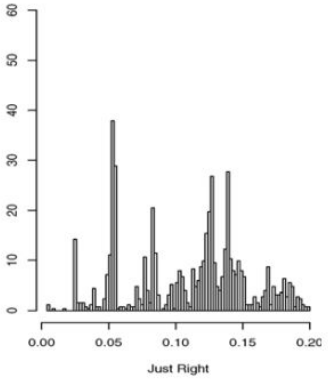
Apunte de Histograma - PyE FIUBA

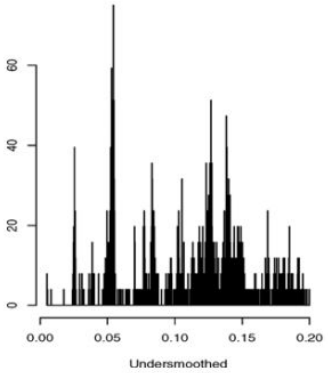
Teorema: Sea x y m fijos, y sea B_n el bin que contiene a x, luego

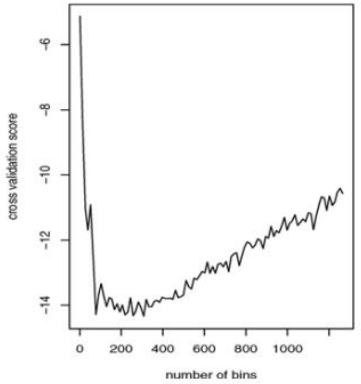
$$\mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h} \qquad \mathbb{V}(\widehat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}.$$

Obs: Al aumentar la cantidad de bins (m), Disminuye el sesgo, pero aumenta la varianza. Acá esta el tradeoff.









Estimación de densidad por kernel

Los histogramas son discontinuos, los estimadores de densidad por kernel (KDE) son una versión más suave y convergen más rápido a la densidad verdadera que el histograma.

Kernels

Se define un kernel como una función K suave tal que:

$K(x) \geq 0$, $\int K(x)dx = 1$, $\int xK(x)dx$=0, y

$\sigma_K^2 = \int x^2 K(x)dx > 0$.

Algunos kernels comunes:

- Epanechinkov: $K(x) = \begin{cases} \frac{3}{4}(1 - x^2/5)/\sqrt{5}, & |x| < 5 \\ 0 & e.o.c. \end{cases}$

  Es óptima en el sentido de error cuadrático medio

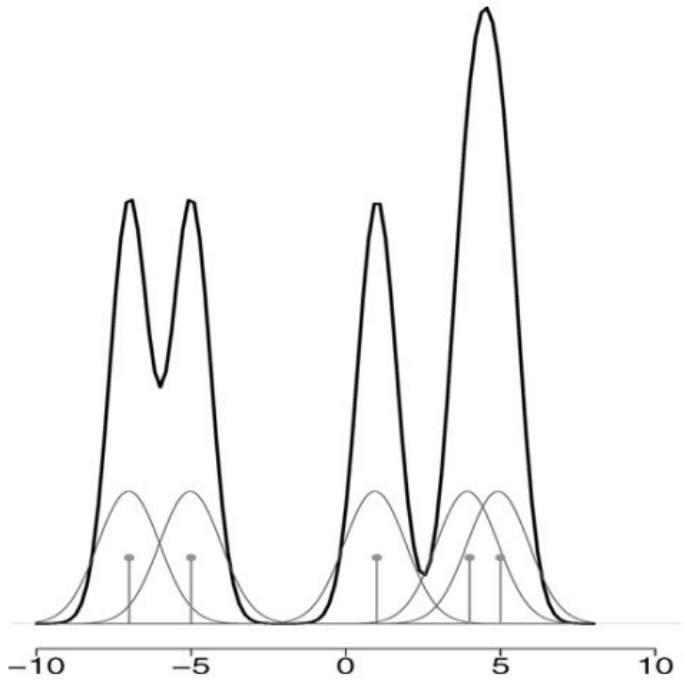
- Gaussiano (simple)

KDE

Def: Dado un kernel K y un número positivo h, llamado ancho de banda, el estimador de densidad por kernel se define como

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

Nuevamente el parámetro h es el que nos controla el tradeoff sesgo-varianza



Bibliografía

- "Notas de Estadística", Graciela Boente y Víctor Yohai, FCEyN, UBA.
- “All of Statistic: A concise Course in Statistical Inference”, Larry Wasserman