Probabilidad y estadística

Clase 5

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.

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

Ejercicio 1

De un experimento en los efectos de un medicamento para la ansiedad, entre otras cosas se midió la diferencia (en segundos) entre el puntaje de un test de memoria antes y después de tomar el medicamento, obteniendo los siguientes resultados:

1.2, 4.6, 4.3, 4.2, –7.9, 7.8, 3.4, 19.8, 25.5, –1.9, 2.1, –0.9, 4.6, 21.1, 1,7

1. Obtener la función de distribución empírica a mano.
2. Utilizar la columna 'Diff' del dataset Islander_data.csv y calcular la func. de distribución empírica usando Python.

Propiedades de la ECDF

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

All of Statistics, Wasserman

Estimación de densidades (*smoothing*)

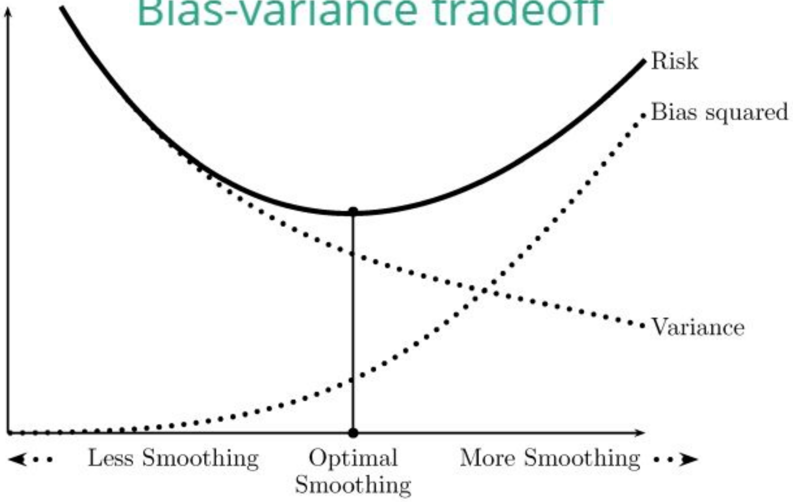
Si deseamos estimar una función de densidad $f(x)$ o una función de regresión $\phi(x) = \mathbb{E}[X|Y = y]$, se deben hacer algunas suposiciones de suavidad.

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

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

$$R(g, \hat{g}_n) = \mathbb{E}\left[\int (g(u) - \hat{g}_n(u))^2 du\right]$$

Bias-variance tradeoff



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

Ejercicio 2

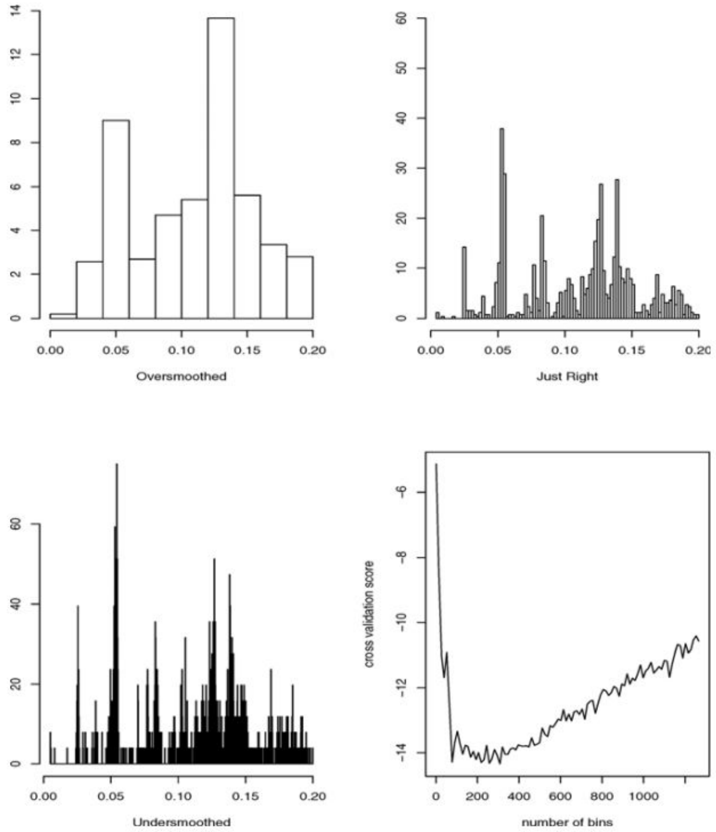
A partir de los datos del ejercicio 1,

1. Calcular a mano, el histograma de 6 bins
2. A partir de los datos del dataset graficar el histograma de la columna 'Diff' utilizando Python

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

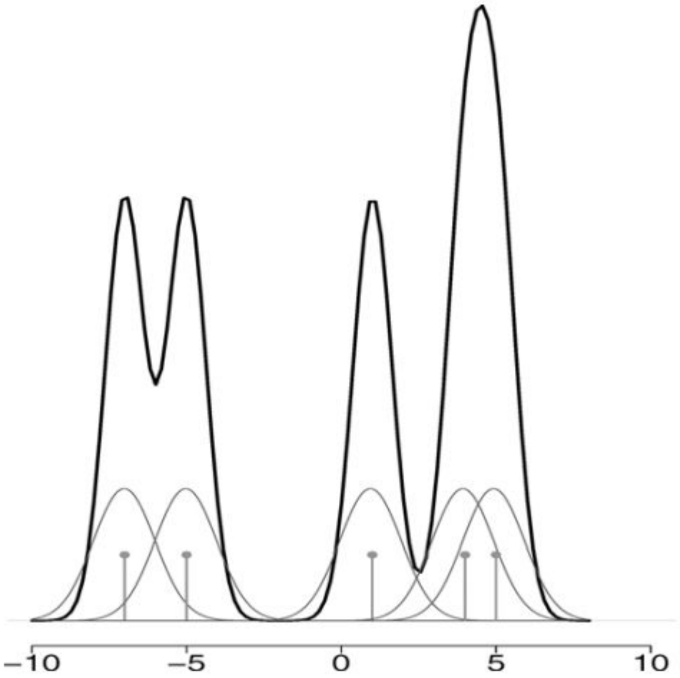
- Gaussiano (simple)

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



Ejercicio 3

A partir de la columna ‘Diff’ del dataset Islander_data estimar la densidad por el método de KDE. Analizar qué ocurre al tomar distintos valores de h.

Intervalos de confianza

Motivación

Hasta ahora habíamos visto estimadores puntuales, que, dada un muestra, nos devuelven un único valor $\hat{\theta}$ que se aproxima al valor verdadero del parámetro deseado θ.

Una forma de obtener información sobre la precisión de la estimación, en el caso de que θ sea unidimensional, es proporcionar un intervalo $[a(X), b(X)]$ de manera que la probabilidad de que dicho intervalo contenga el verdadero valor θ sea alta, por ejemplo, 0.95.

Región de confianza

Def: Dada una m.a. $\underline{X}$ con distribución perteneciente a una familia $F_\theta(x)$, con $\theta \in \Theta$, una región de confianza $S(\underline{X})$ para θ con nivel de confianza $1-\alpha$ será un conjunto tal que

$$\mathbb{P}(\theta \in S(\underline{X})) = 1-\alpha. \ (*)$$

Obs: θ **no** es aleatorio, lo aleatorio es (*) es $S(\underline{X})$.

Obs: Si $S(\underline{X}) = (a(\underline{X}), b(\underline{X}))$ diremos que es un intervalo de confianza.

Si $S(\underline{X}) = (\min(\Theta), b(\underline{X}))$ diremos que es una cota superior.

Si $S(\underline{X}) = (a(\underline{X}), \max(\Theta))$ diremos que es una cota inferior.

Juguemos un poquito

Usemos la siguiente api para entender mejor qué es un IC

Método del pivote

Teorema: Sea $\underline{X}$ una muestra aleatoria con distribución perteneciente a una familia $F_\theta(x)$, con $\theta \in \Theta$, y sea $U = g(\underline{X}, \theta)$ una variable cuya distribución **no** depende de θ. Sean a y b tales que $\mathbb{P}(a \leq U \leq b) = 1 - \alpha$. Luego,

$$S(\underline{X}) = \{\theta : a < g(\underline{X}, \theta) \leq b\}$$

es una región de confianza para θ. A U se lo llama pivote.

Ejercicio 4

Sea $\underline{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de tamaño n de una población con distribución uniforme en el intervalo $(0, \theta)$. Hallar una cota inferior del 95% para . θ

Suponer n=20 y θ =3, simular la muestra y obtener el valor de la cota

Algunos resultados importantes

Teorema: Sea $\underline{X} = X_1, \ldots, X_n$ una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$

- $Z = \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \sim \mathcal{N}(0,1)$
- $W = \sum_{i=1}^{n} \frac{(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$
- V y W son independientes
- Si $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, U = \sqrt{n}\frac{(\bar{X}-\mu)}{S} \sim t_{n-1}$

Obs: en general vale que si $X \sim \mathcal{N}(0,1)$ y $Y \sim \chi^2_n$, con X e Y independientes vale que $\frac{X}{\sqrt{Y/n}} \sim t_n$

Algunos pivotes para variables normales

Dada $\underline{X}_n$ una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$. Definimos algunos pivotes:

- Para la media con σ^2 conocida: $U(\underline{X}, \mu) = \frac{X-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$
- Para la media con σ^2 desconocido: $U(\underline{X}, \mu) = \frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}$
- Para el desvío con media conocida $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma} \sim \chi^2_n.$
- Para el desvío con media desconocida $\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma} \sim \chi^2_{n-1}.$

Dada también $\underline{Y}_m$ una m.a. de una distribución $\mathcal{N}(\lambda, \sigma^2)$

- Comparación de medias con varianzas conocida e iguales: $\frac{\bar{X}-\bar{Y}-(\mu-\lambda)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}$
- Comparación de medias con varianzas conocida e iguales: $\frac{\bar{X}-\bar{Y}-\Delta}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$, con

$$S_p^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{m+n-2}$$

Algunos pivotes para variables normales

Dada $\underline{X}_n$ una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$ definimos algunos pivotes:

- Para la media con varianza conocida: $U(\underline{X}, \mu) = \frac{(\overline{X}-\mu)}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$
- Para la media con varianza desconocida: $U(\underline{X}, \mu) = \frac{(\overline{X}-\mu)}{S}\sqrt{n} \sim t_{n-1}$
- Para el desvío con media conocida: $U(\underline{X}, \sigma) = \frac{\sum_{i-1}^{n}(X_i-\mu)^2}{\sigma} \sim \chi^2_n$
- Para el desvío con media desconocida: $U(\underline{X}, \sigma) = \frac{\sum_{i-1}^{n}(X_i-\overline{X})^2}{\sigma} \sim \chi^2_{n-1}$

Dada también $\underline{Y}_m$ una m.a. de una distribución $\mathcal{N}(\lambda, \sigma^2)$ y sea :

- Comparación de medias con varianzas conocidas: $U(\underline{X}, \Delta) = \frac{\overline{X}-\overline{Y}-\Delta}{\sqrt{\frac{\sigma^2_X}{n}+\frac{\sigma^2_Y}{m}}} \sim \mathcal{N}(0,1)$
- Comparación de medias con varianzas desconocidas e iguales:

$$U(\underline{X}, \Delta) = \frac{\overline{X}-\overline{Y}-\Delta}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$, con $$S_p^2 = \frac{(m-1)S_X^2+(n-1)S_Y^2}{n+m-2}$$

Ejercicio 5

Dada una muestra aleatoria $\underline{X} = (X_1, \ldots, X_n)$de una población con distribución normal con media y varianza desconocidas, hallar el intervalo de confianza de nivel 0.99 para la media de la población.

Suponer n=50, $\mu = 2, \sigma = 3$, simular la muestra y calcular el IC resultante de la misma.

Regiones de confianza asintóticas

Def: Sea $\underline{X}_n = X_1, \ldots, X_n$ una m.a de una población con distribución perteneciente a la flía. $F_\theta(x)$, con $\theta \in \Theta$. Se dice que $S_n(\underline{X}_n)$ es una sucesión de regiones de confianza de nivel asintótico $1-\alpha$ si:

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in S_n(\underline{X}_n)) = 1-\alpha$$

Teorema: Sea $\underline{X}_n$ una m.a. de una población con distribución $F_\theta(x)$, con $\theta \in \Theta$. Supongamos que para cada n se tiene $U_n = g(\underline{X}_n, \theta)$ que converge en distribución a U, donde U es una v.a. cuya distribución no depende de θ. Entonces si a y b son tales que $\mathbb{P}(a < U < b) = 1-\alpha$ se tiene que $S_n(\underline{X}_n) = \{\theta : a < U_n < b\}$ es una región de confianza de nivel asintótico $1-\alpha$ para θ.

Ejercicio 6

Se arroja 50 veces una moneda con probabilidad p de salir cara. Hallar un intervalo de confianza asintótico de nivel 0.95 para p basado en la observación $x=50$.

IC para la media de una población desconocida

En general, dada una m.a $\underline{X}_n$ de una población desconocida, una buena forma de aproximarse a la media de dicha población es considerar el promedio de las muestras ($\bar{X}_n$).

Por TCL, sabemos que $\bar{X}_n$ tiende en distribución a una v.a. normal. En particular,

$$\frac{\bar{X}_n - \mathbb{E}[X]}{\sqrt{var(X)/n}} \overset{(a)}{\sim} \mathcal{N}(0,1)$$

Se puede probar que si se desconoce también la varianza de la población (que es lo más común) vale que

$$\frac{\bar{X}_n - \mathbb{E}[X]}{S/\sqrt{n}} \overset{(a)}{\sim} \mathcal{N}(0,1)$$

Ejercicio 7

De un experimento en los efectos de un medicamento para la ansiedad se midió el puntaje en un test de memoria antes y después de tomar el medicamento. A partir de los datos que se encuentran en el archivo Islander_data.csv hallar un IC para la media del tiempo de respuesta después de consumir el medicamento.

Bibliografía

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