

Probabilidad y estadística

Clase 5

Estimación no paramétrica

Función de distribución empírica

Tenemos tal que

$X_i, \quad i=1, \dots, n$ son iid.

Función de distribución empírica (ECDF):

Es una aproximación a la función de distribución , que pone peso $1/n$ a cada observación

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

$$F_x(x) = P(X \leq x)$$

Propiedades de la ECDF

Para cada $x \in \mathbb{R}$,

$$\mathbb{E}(\widehat{F}_n(x)) = F(x),$$

$$\mathbb{V}(\widehat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n},$$

$$\text{MSE} = \frac{F(x)(1 - F(x))}{n} \rightarrow 0,$$

$$\left[\widehat{F}_n(x) \xrightarrow[n \rightarrow \infty]{\text{P}} F(x) \right]$$

Ejercicio 1

Usemos el [Advertising Sales Dataset](#). Allí se presentan valores del presupuesto asignado (en 1000\$) en distintos medios (TV, radio, diarios) y las ventas asociadas.

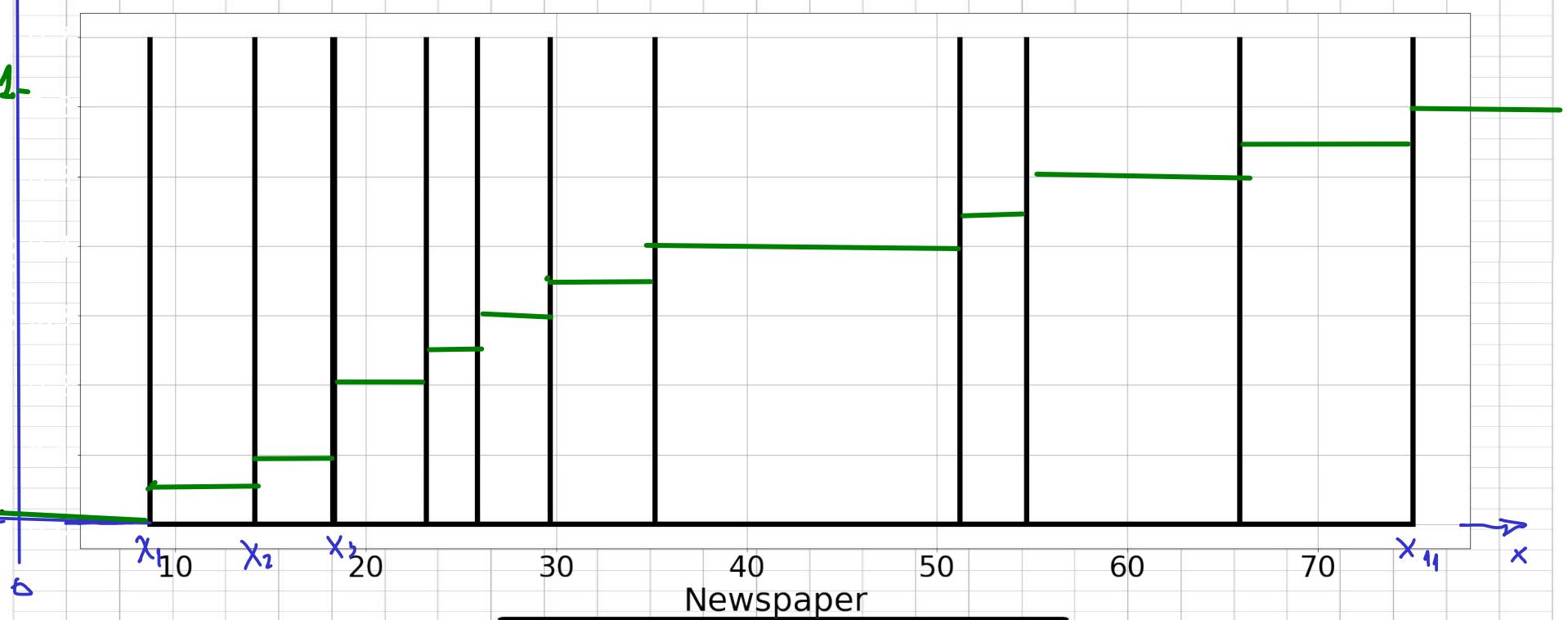
1. A partir de la muestra 8.7, 14.2, 18.3, 18.4, 23.2, 25.9, 29.7, 35.2
51.2, 54.7, 65.9, 75 obtener la función de distribución empírica a mano.
2. Utilizar la columna “Newspaper” del archivo “advertising.csv” y calcular la func. de distribución empírica usando Python.

A partir de la muestra 8.7, 14.2, 18.3, 18.4, 23.2, 25.9, 29.7, 35.2

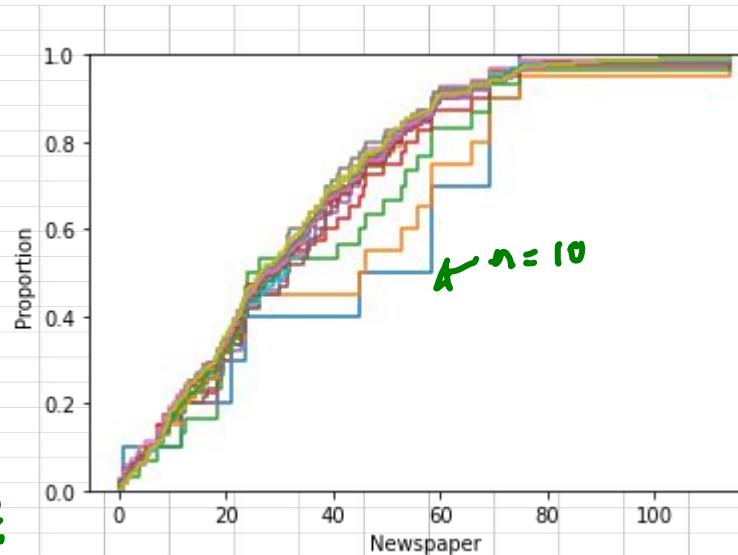
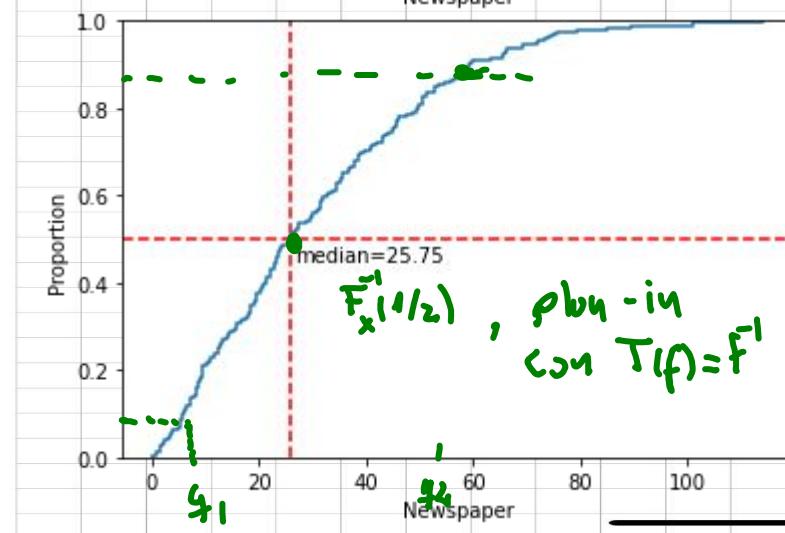
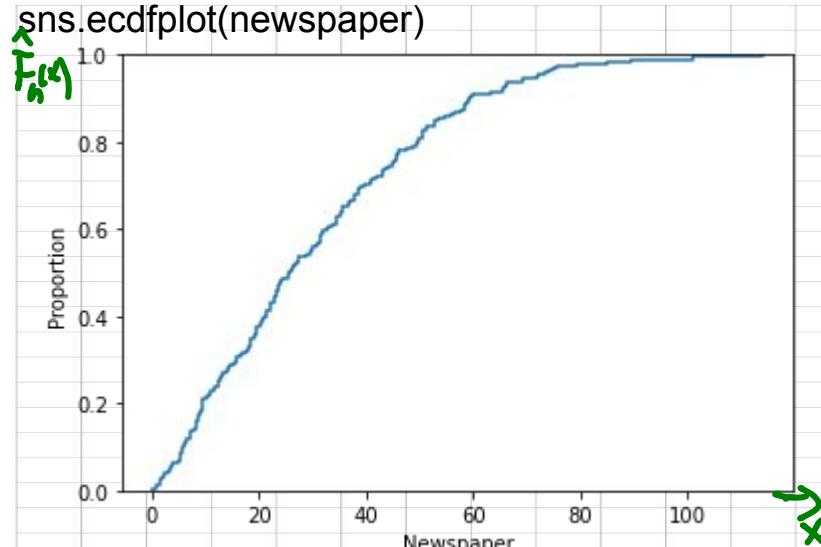
51.2, 54.7, 65.9, 75 obtener la función de distribución empírica a mano.

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$$

elector.0



sns.ecdfplot(newspaper)



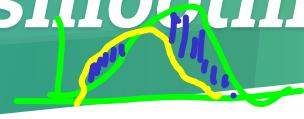
Estimador “plug-in”: $\hat{\theta} = T(\hat{F}_n)$

obs:

$$\left[\hat{F}_n^{-1} = \inf\{x : \hat{F}_n \geq p\} \right]$$

mediana $\hat{F}_n^{-1}(1/2)$, quantil q $\hat{F}_n^{-1}(q)$

Estimación de densidades (smoothing)



$$x \sim F_\theta(x)$$

$$\text{ECM} = E[(\theta - \hat{\theta})^2] = \text{Bias}^2 + \text{Var}$$

A la hora de estimar **funciones** de densidad, queremos tener una medida de cuán buena es la estimación. (Equivalente al ECM para parámetros)

, o **ECMI**.

Para densidades vamos a definir el **riesgo**:

densidad de X

$$R(g, \hat{g}_n) = \mathbb{E}\left[\int_{-\infty}^{\infty} \{g(x) - \hat{g}_n(x)\}^2 dx\right]$$

modelos para mi densidad

$$= \int_{-\infty}^{\infty} b^2(x)dx + \int_{-\infty}^{\infty} v(x)dx$$

$\hat{g}_n(x)$ es aleatorio

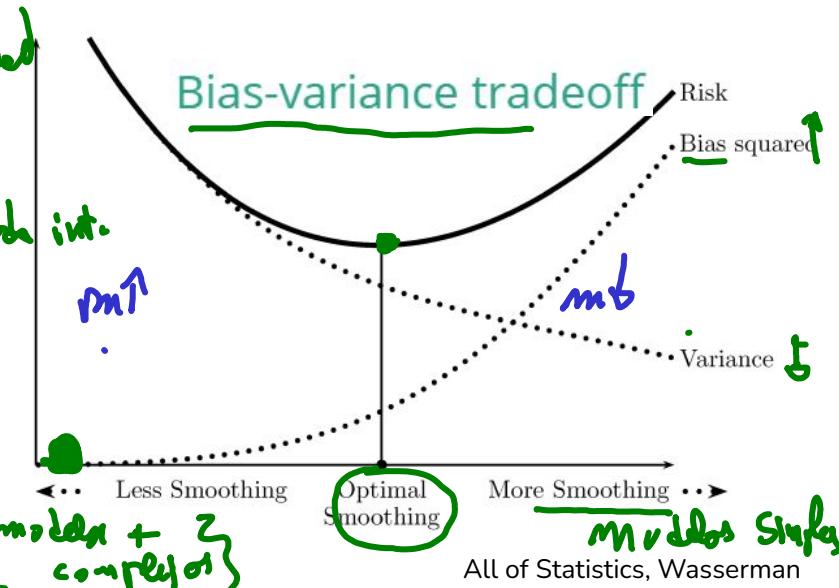
$b(x)$ bias

$v(x)$ var.

$$b(x) = \mathbb{E}[\hat{g}_n(x)] - g(x)$$

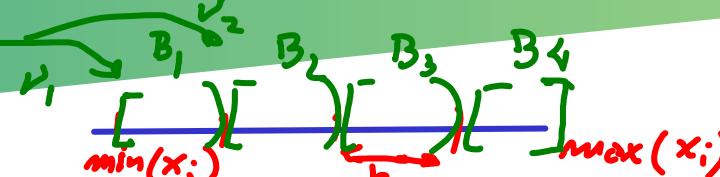
$$v(x) = \mathbb{E}[\{\hat{g}_n(x) - \mathbb{E}[\hat{g}_n(x)]\}^2]$$

"Overfitting" {mod. + comp.}

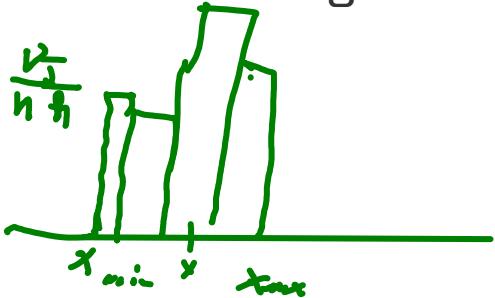


$x_i : i=1 \dots n$ son iid (muestra) → Normalizar los datos al $[0, 1]$

Histogramas



1. Se toman los valores máximo y mínimo y se divide el intervalo en m sub-intervalos de longitud h . A cada subintervalo lo llamaremos B_j .
2. Se cuenta la cantidad de observaciones que caen en cada B_j :
 $\nu_j = \sum_{i=1}^n 1\{X_i \in B_j\}$.
3. Normalizamos dividiendo por la cantidad total de muestras n , y por la longitud del subintervalo h .



$$\hat{f}_n(x) = \left(\frac{1}{nh} \sum_{j=1}^m \nu_j 1\{x \in B_j\} \right) \checkmark$$

$$\hat{f}_n(x) = \frac{1}{h} \sum_{j=1}^m \hat{p}_j 1\{x \in B_j\} \quad \text{donde} \quad \boxed{\hat{p}_j = \nu_j/n}$$

$$E[\hat{f}_n(x)] = \frac{1}{n} E \hat{P}_{J_0} = \frac{P_{J_0}}{n} = \frac{\int_{B_{J_0}} f_X(x) dx}{n} \stackrel{\text{TVM}}{\approx} \frac{\int_{B_{J_0}} f_X(x) dx}{n} P_{J_0} = \int_{B_{J_0}} f_X(x) dx$$

Ejercicio 2

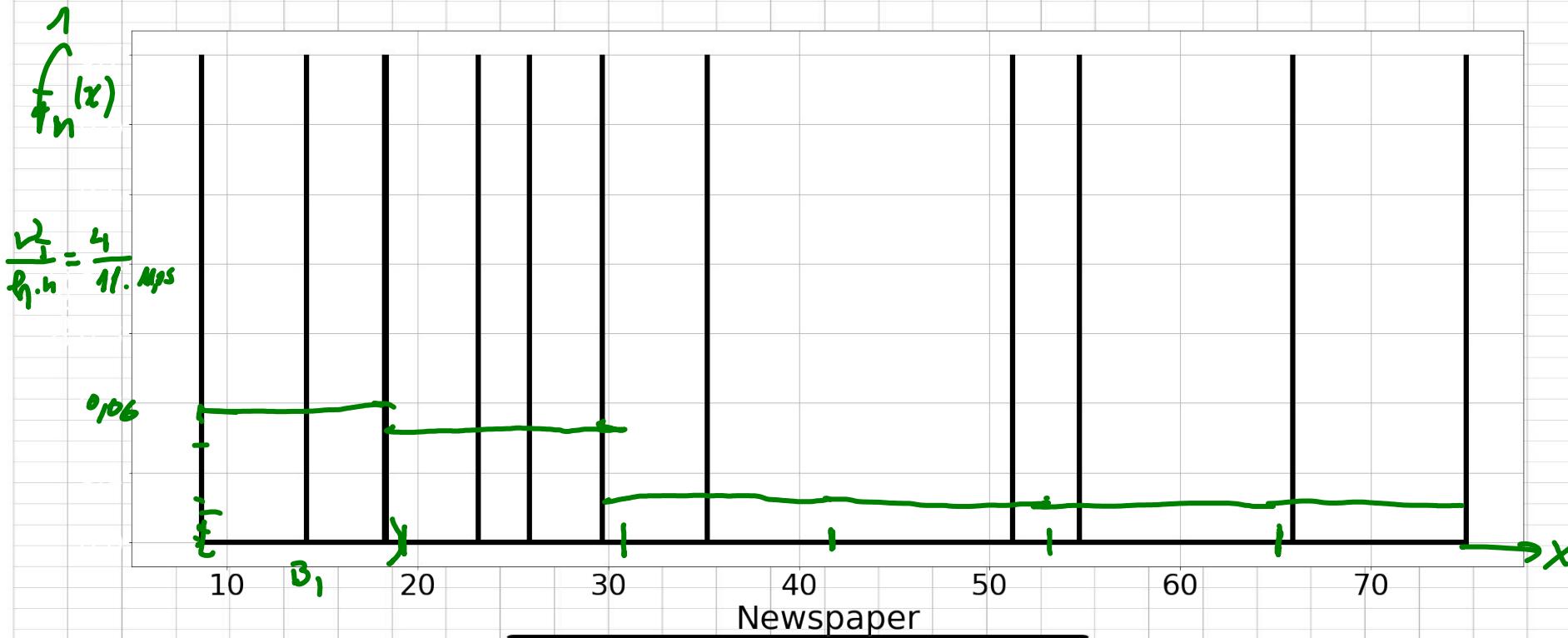
A partir de los datos del ejercicio 1,

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

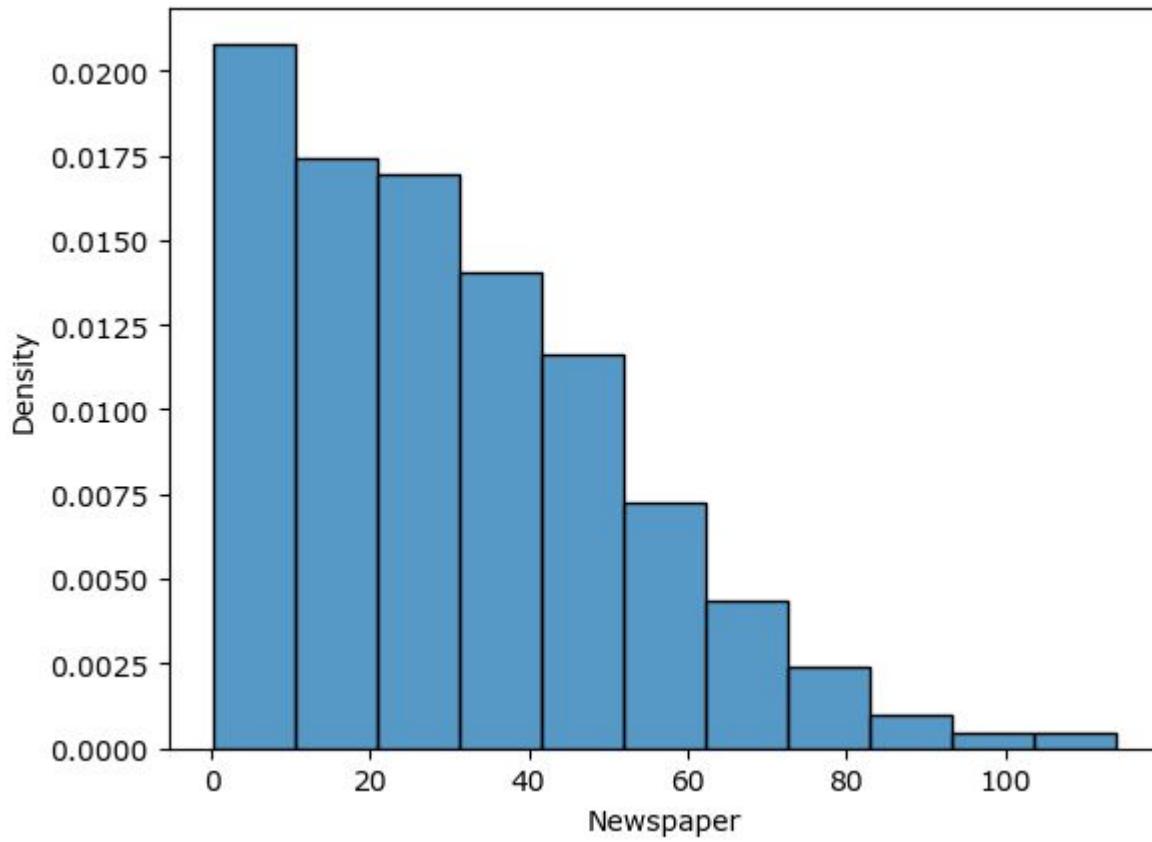
$$[8.7, 14.2, 18.3, 18.4] \quad [23.2, 25.9, 29.7, 35.2] \quad [51.2, 54.7, 65.9, 75]$$

$$h = \frac{x_{\max} - x_{\min}}{6} = \underline{11.05}$$

$$\frac{8.7 + 2.4}{3} = \underline{30.8}$$



```
sns.histplot(newspaper, stat='density')
```

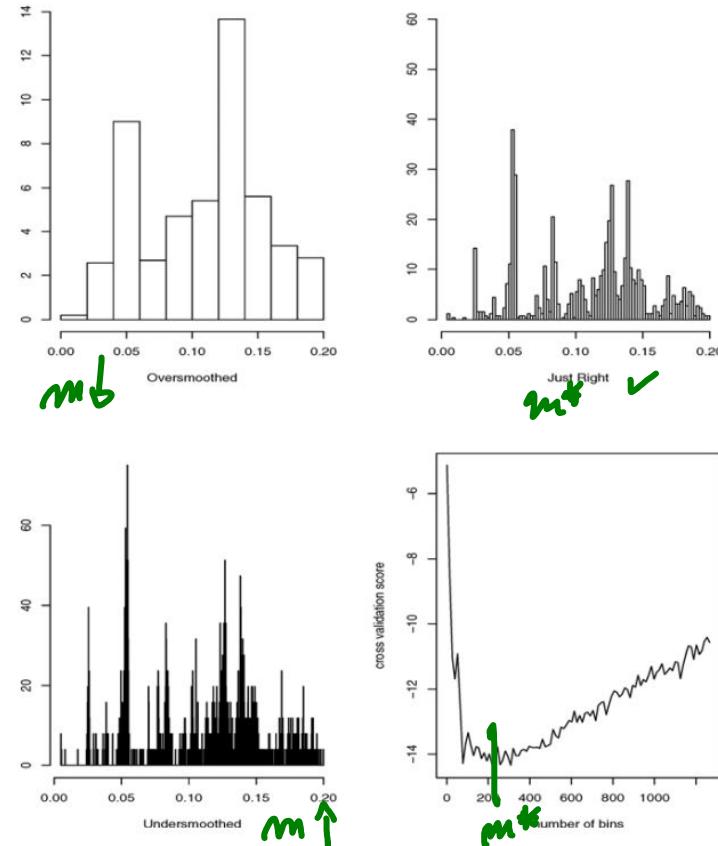


Propiedades del histograma

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

$$\mathbb{E}(\hat{f}_n(x)) = \frac{p_j}{h} \xrightarrow{n \rightarrow \infty} f(x), \quad \text{Var}(\hat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}.$$

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

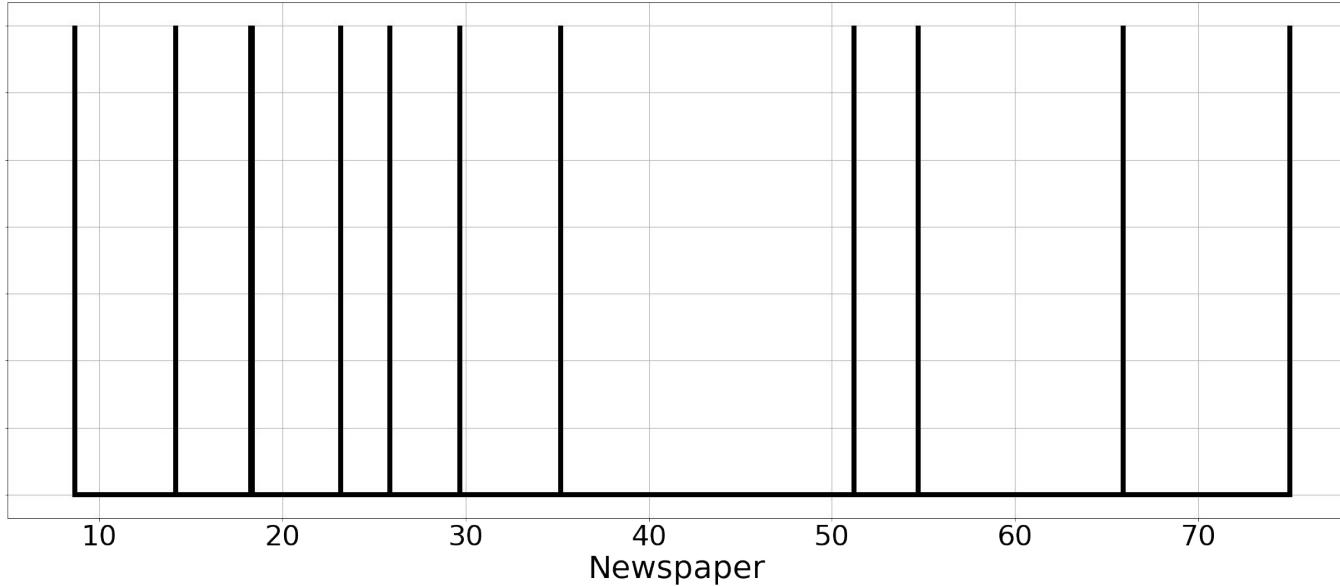
Existen los **estimadores de densidad por kernel (KDE)**, que son más suaves y convergen más rápido a la verdadera densidad de los datos.

Estos estimadores asignan un peso a cada muestra que se “desparrama” a los puntos vecinos



Kernel density estimation

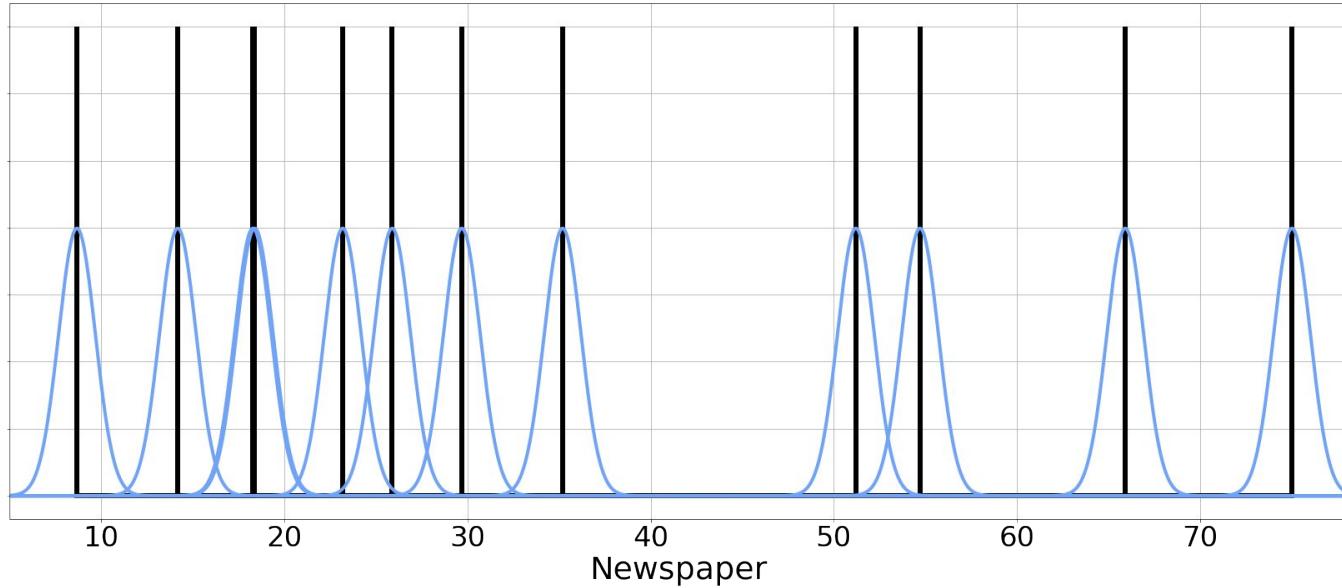
Primero:
marcamos las
observaciones en
el eje x



Kernel density estimation

Segundo:

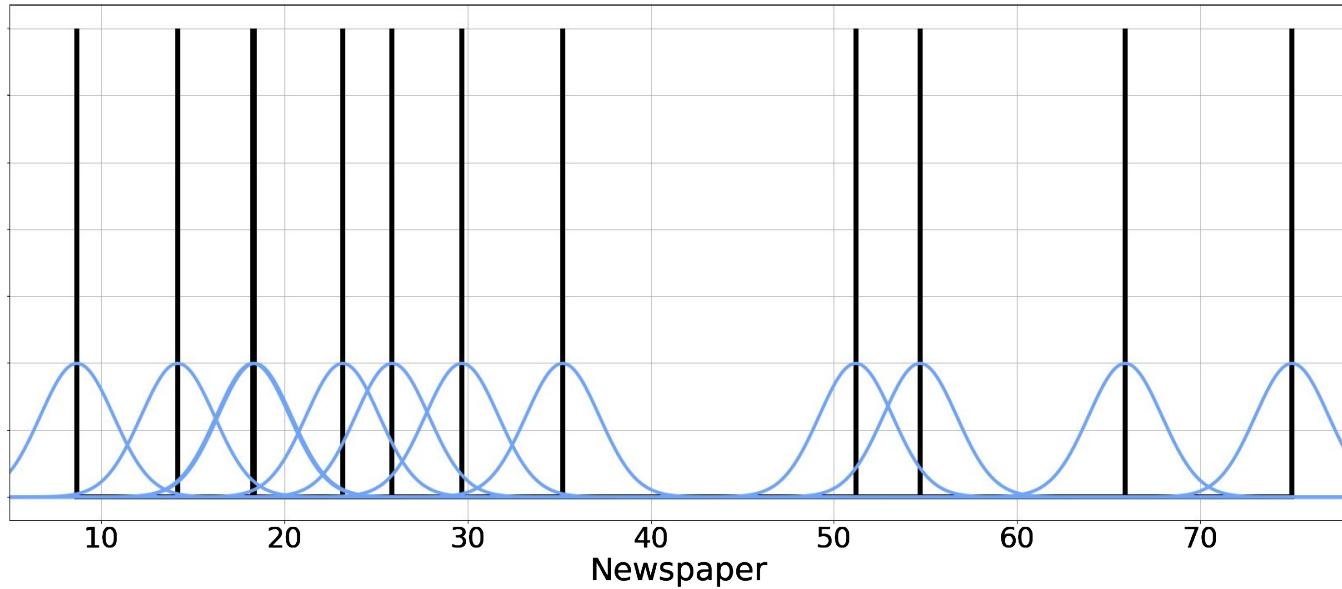
Montamos una función (kernel) sobre cada muestra



Kernel density estimation

Segundo:

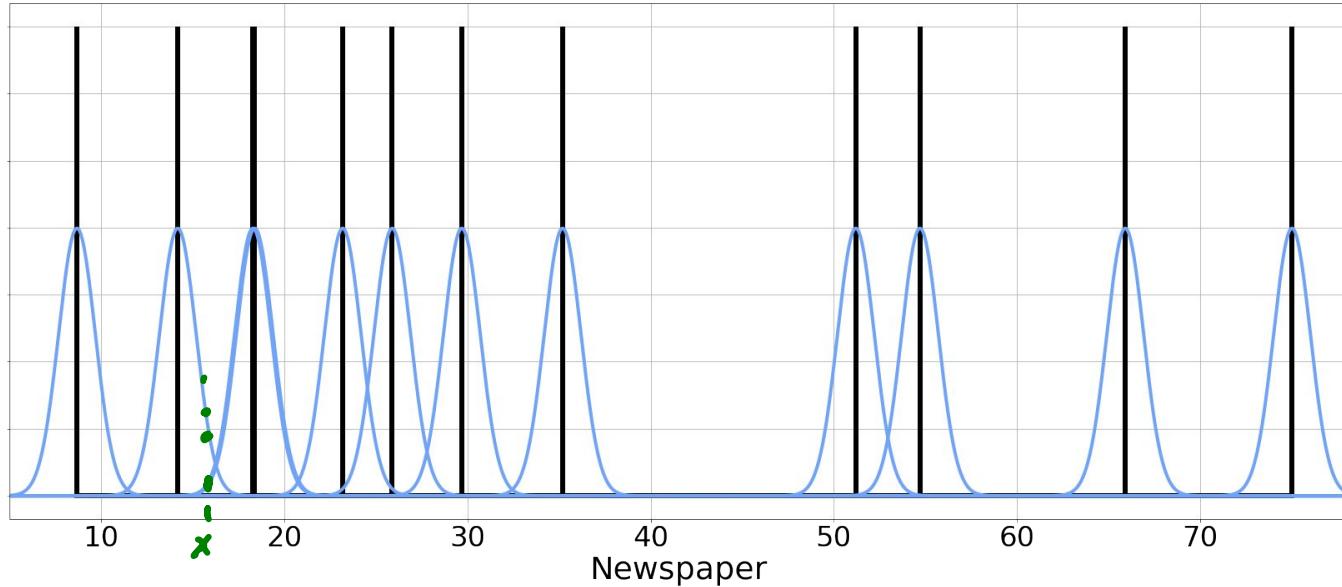
Montamos una función (kernel) sobre cada muestra



Kernel density estimation

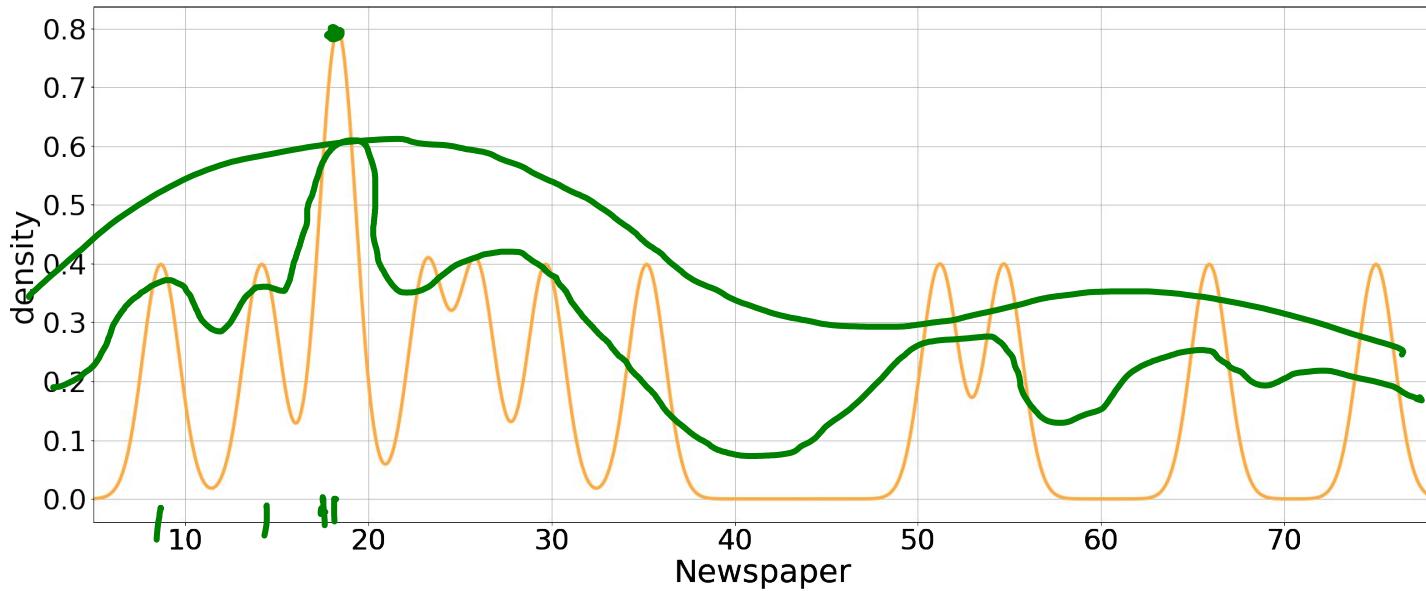
Segundo:

Montamos una función (kernel) sobre cada muestra



Kernel density estimation

Tercero: dividimos todo por n y sumamos las curvas



Kernels

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

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

✓ $\sigma_K^2 = \int x^2 K(x)dx > 0.$ ✓ $\bar{x} = 0$

Algunos kernels comunes:

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

Es óptima en el sentido de error cuadrático medio ✓

- Gaussiano (simple) ✓

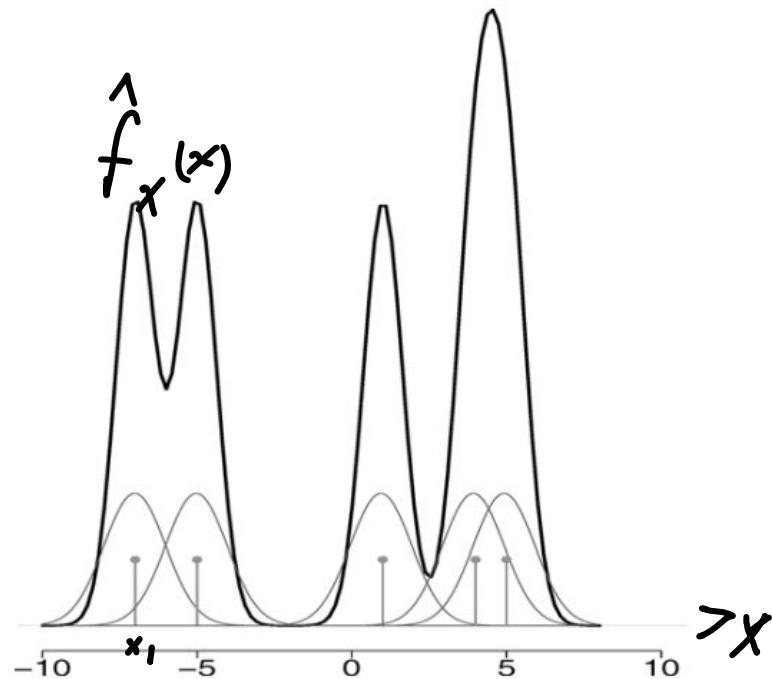
KDE

$$K \sim f_X(x) \text{ con } X \sim \mathcal{N}(0,1)$$

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

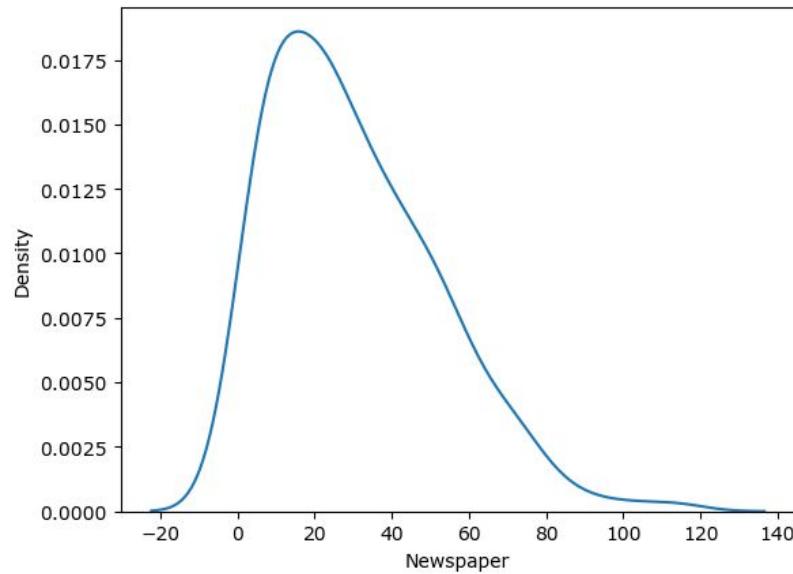
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} H\left(\frac{x-X_i}{h}\right)$$

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



Ejercicio 3

A partir de la columna ‘Newspaper’ del dataset estimar la densidad por el método de KDE.



Intervalos de confianza

Motivación

$X \sim F_\theta \Leftrightarrow$ es una distri. parámetrica.

$\underline{x}_n = \{x_i\}_{i=1}^n$ "la muestra aleatoria" x_i : son iid.
 x_i iid $i=1 \dots n$ $x_i \sim \bar{\theta}$

Hasta ahora habíamos visto estimadores puntuales, que, dada una 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

$$x \sim N(\theta, \sigma^2), \quad \sigma^2 > 0 \\ \Theta = \{\theta : \sigma^2 > 0\}$$

iid

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

$$\underline{\mathbb{P}(\theta \in S(\underline{X}))} = 1 - \alpha. (*) \quad (\text{a en chequero})$$

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

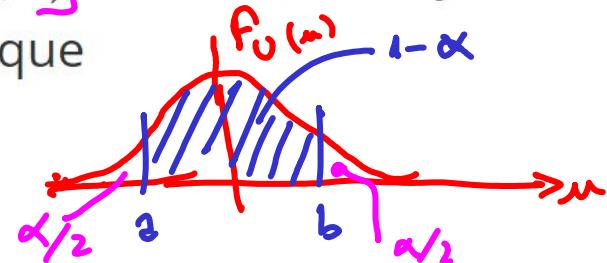
<http://rossmanchance.com/applets/2021/confsim/ConfSim.html>

Método del pivote

$$\left\{ \begin{array}{l} X_i \sim N(\mu, \sigma^2) \quad \text{"estandarizar"} \\ \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad U = \frac{\bar{X} - \mu}{\sigma} \sim N(0,1) \\ \text{No depende de } \mu \quad n: \text{de } \sigma \end{array} \right.$$

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

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$N(0,1) \sim U = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}-\mu}{\sigma} \sqrt{n}$$

Sea $\underline{X}_n = (X_1, \dots, X_n)$ una muestra aleatoria de tamaño n de una población con distribución normal de media μ ? y varianza 4. Hallar una cota inferior del 95% para μ .

Suponer n=20 y $\mu=3$, simular la muestra y obtener el valor de la cota.

$$U = \frac{\bar{X}-\mu}{\sigma} \sqrt{n} \sim N(0,1) , \quad U \text{ es función decreciente de } \mu$$

$$P(U < a) = 0,95$$



$$F_U(x) = P(U \leq x)$$

$$U < F_U^{-1}(0,95) = \{_{0,95} \approx 1,64$$

$$U < 1,64$$

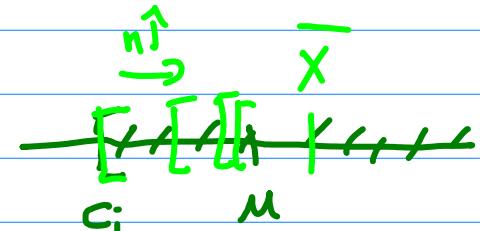
$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sqrt{n} < 1,64$$

$$\boxed{\mu > \bar{X} - \frac{2 \cdot 1,64}{\sqrt{n}}}$$

es cota inf con prob 0,95.

$$P\left(\mu > \bar{X} - \frac{2 \cdot 1,64}{\sqrt{n}}\right) = 0,95.$$

$\underbrace{c_i}_{\substack{\rightarrow 0 \\ n \rightarrow \infty}}$



z) $\bar{X} \sim N(3, 4)$ con $n=20$.

calcular \bar{X} y calcular lo ci $\bar{X} - \frac{2 \cdot 1,64}{\sqrt{n}}$.

Algunos resultados importantes

MLE para σ^2 es $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
es sesgado.

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

RC para μ [$Z = \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma}$ ~ $\mathcal{N}(0, 1)$]
constante σ

RC para σ^2 [$W = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$]

X^2 con $X \sim \mathcal{N}(0, 1)$
es una χ^2_1 .

V y W son independientes

Est. sesgado de la varianza
Si $[S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2]$ [$U = \sqrt{n} \frac{(\bar{X} - \mu)}{\underline{S}} \sim t_{n-1}$] RC para μ

{ 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: $U(\underline{X}, \mu) = \frac{\underline{X} - \mu}{\sigma} \sqrt{n} \sim \mathcal{N}(0, 1)$ ✓

Para la media con σ^2 desconocido: $U(\underline{X}, \mu) = \frac{\underline{X} - \mu}{\underline{S}} \sqrt{n} \sim t_{n-1}$

Para el desvío con media conocida $\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma} \sim \chi_n^2$.

Para el desvío con media desconocida $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma} \sim \chi_{n-1}^2$.

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: $U(\underline{X}, \mu) = \frac{(\bar{X} - \mu)}{\sigma} \sqrt{n} \sim \mathcal{N}(0, 1)$ •
- Para la media con varianza desconocida: $U(\underline{X}, \mu) = \frac{(\bar{X} - \mu)}{\underline{S}} \sqrt{n} \sim t_{n-1}$ $n \rightarrow \infty$ $\xrightarrow{n \rightarrow \infty} \mathcal{N}(0, 1)$
- Para el desvío con media conocida: $U(\underline{X}, \sigma) = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$
- Para el desvío con media desconocida: $U(\underline{X}, \sigma) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$

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

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

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

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

Ejercicio 5

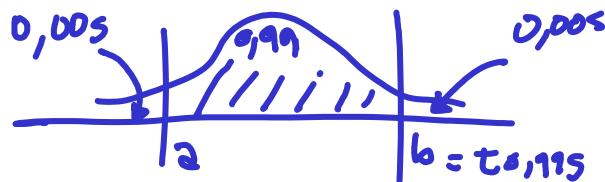
Dada una muestra aleatoria $\underline{X} = (X_1, \dots, X_n)$ de una población con distribución normal con media y varianza desconocidas, hallar el intervalo de confianza de nivel 0.99 para la media de la población.

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

$$U = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim t_{n-1}$$

$$P(a < U < b) = 0,99 \quad \text{usando que } U \sim t_{n-1}$$

$$F_U^{-1}(0,005) < U < F_U^{-1}(0,995) \Rightarrow t_{0,995} \approx z_{0,05} = -t_{0,995}$$



$$S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})$$

$$P\left(-t_{995} < \frac{\bar{x} - \mu}{S} \sqrt{n} < t_{995}\right) = 0,99$$

$$-\bar{x} - t_{995} \frac{S}{\sqrt{n}} < -\mu < t_{995} \frac{S}{\sqrt{n}} - \bar{x}$$

$$\bar{x} + t_{995} \frac{S}{\sqrt{n}} > \mu > \bar{x} - t_{995} \frac{S}{\sqrt{n}}$$

$$-t_{995} \frac{S}{\sqrt{n}} + \bar{x} < \mu < t_{995} \frac{S}{\sqrt{n}} + \bar{x} \quad \checkmark$$

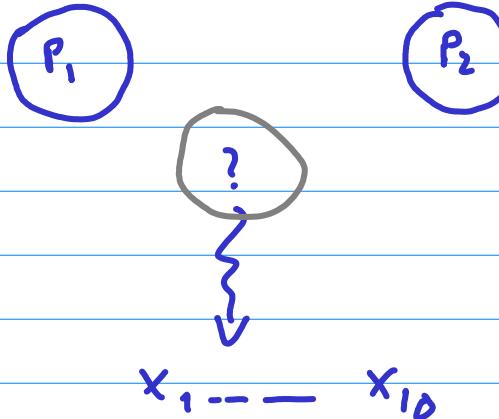
$$\left\{ P(\cdot) = f(p) \stackrel{\text{ppio de inv.}}{\Rightarrow} \hat{P}_{\text{MLE}}(\cdot) = f(\hat{p}_{\text{MLE}}) \right\}$$



$$\text{Es } \Rightarrow \left(\frac{z}{s} - \varepsilon, \frac{z}{s} + \varepsilon \right) \subset \Theta?$$

No es abierto

1. Una moneda tiene una probabilidad de cara p , $p \in \{2/5; 4/5\}$. En 10 lanzamientos de la moneda se observaron exactamente 3 caras. Estimar por máxima verosimilitud la probabilidad de que en otros tres lanzamientos se observe exactamente una cara.

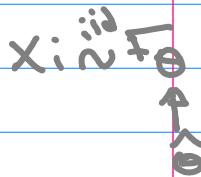


$$\text{Si } \Theta \text{ es un } \underline{\text{abierto}}, \quad \hat{p} : \frac{d \mathcal{L}}{d p} = 0$$

$$\Theta \text{ es abierto si } \forall \theta \in \Theta \quad \exists \varepsilon > 0 : (\theta - \varepsilon, \theta + \varepsilon) \subset \Theta. \\ \equiv B(\theta) \subset \Theta$$

\hat{p} por MLE,

$$\mathcal{L}(\hat{p}) = \max_{p \in \Theta} \mathcal{L}(p)$$





Regiones de confianza asintóticas

Def: Sea $\underline{X}_n = X_1, \dots, X_n$ 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:

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

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,

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

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

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

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