Probabilidad y estadística

Clase 5

Estimación no paramétrica

Función de distribución empírica

Tenemos tal que

Función de distribución empírica (ECDF):
Basándose en una muestra de tamaño n, aproxima la función de distribución poblacional, poniendo una masa de probabilidad puntual de 1/n en cada observación:

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

Propiedades de la ECDF

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

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

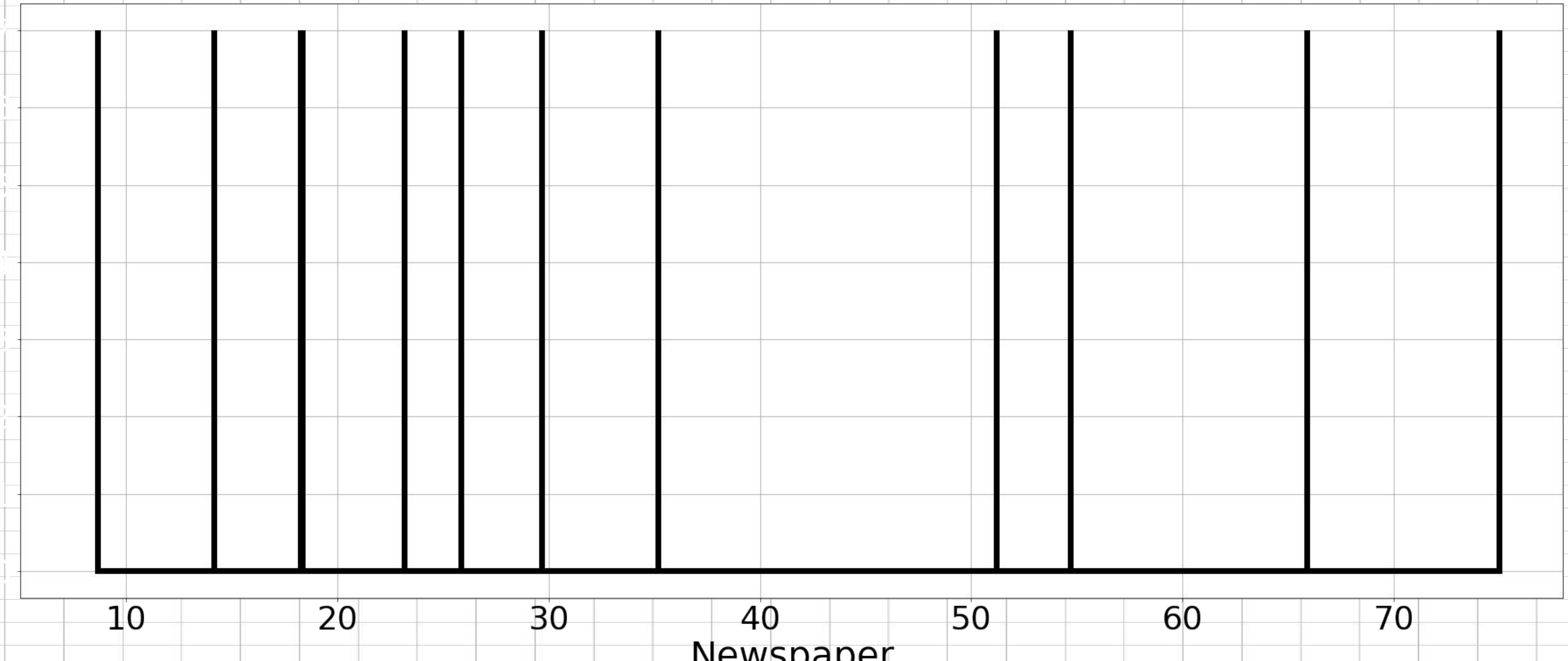
All of Statistics, Wasserman

Ejercicio 1

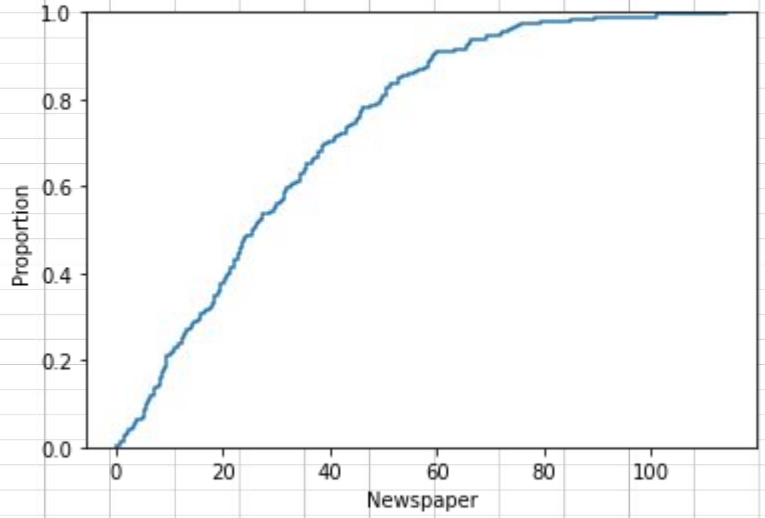
Usemos el Advertising Sales Dataset. Allí se presentan valores del presupuesto asignado (en 1000$) en distintos medios (TV, radio, diarios) y las ventas asociadas.

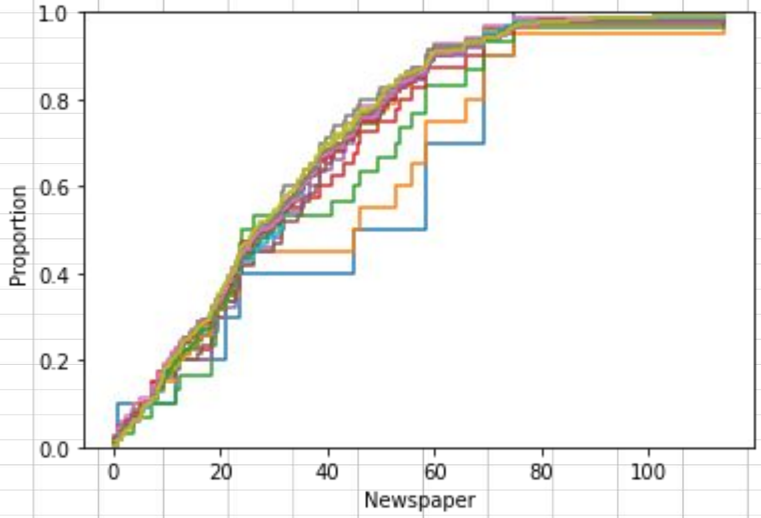
1. A partir de la muestra 8.7, 14.2, 18.3, 18.4, 23.2, 25.9, 29.7, 35.2 51.2, 54.7, 65.9, 75 obtener la función de distribución empírica a mano.
2. Utilizar la columna “Newspaper” del archivo “advertising.csv” y calcular la func. de distribución empírica usando Python.

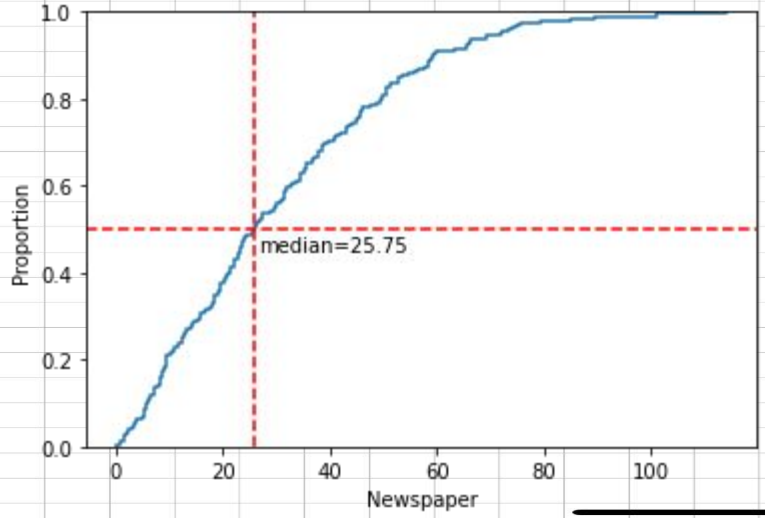
A partir de la muestra 8.7, 14.2, 18.3, 18.4, 23.2, 25.9, 29.7, 35.2
51.2, 54.7, 65.9, 75 obtener la función de distribución empírica a
mano.



sns.ecdfplot(newspaper)







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

obs:

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

Estimación de densidades (*smoothing*)

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

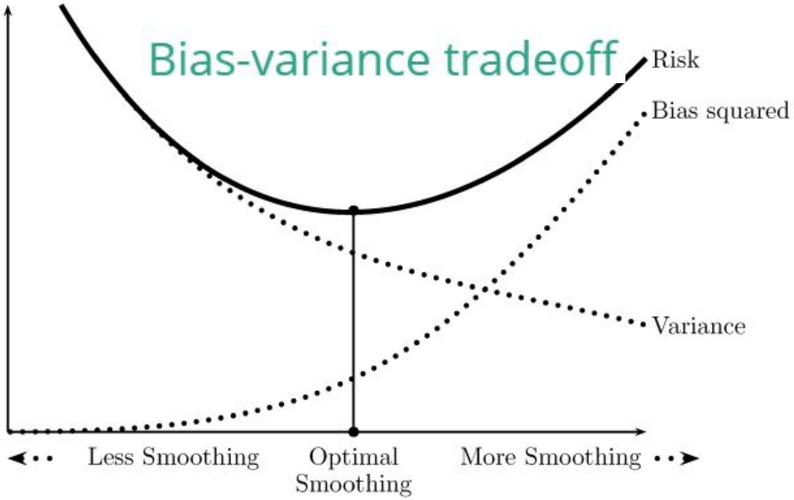
Para densidades vamos a definir el riesgo:

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

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

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

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



All of Statistics, Wasserman

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.

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

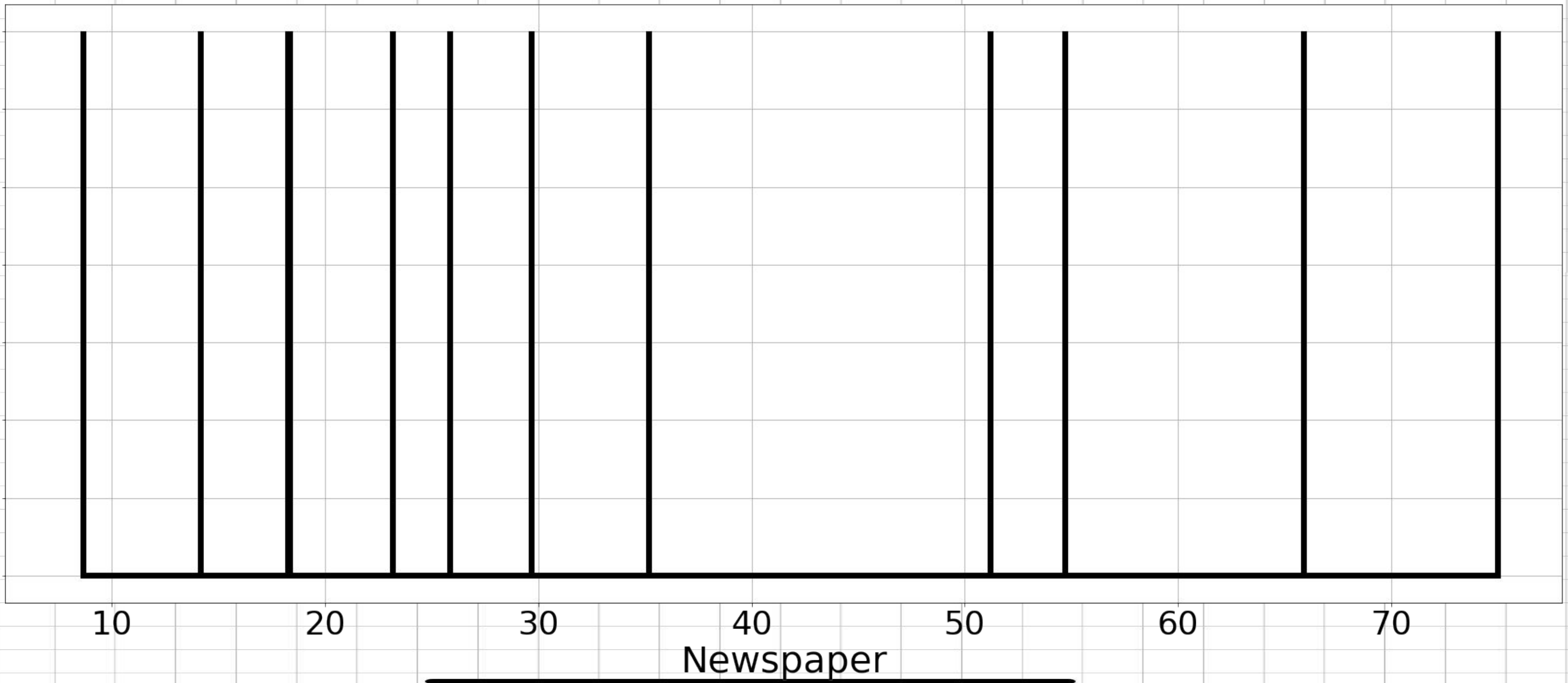
$$\hat{f}_n(x) = \frac{1}{h} \sum_{j=1}^{m} \hat{p}_j 1\{x \in B_j\}$$ donde $$\hat{p}_j = \nu_j / n$$

Ejercicio 2

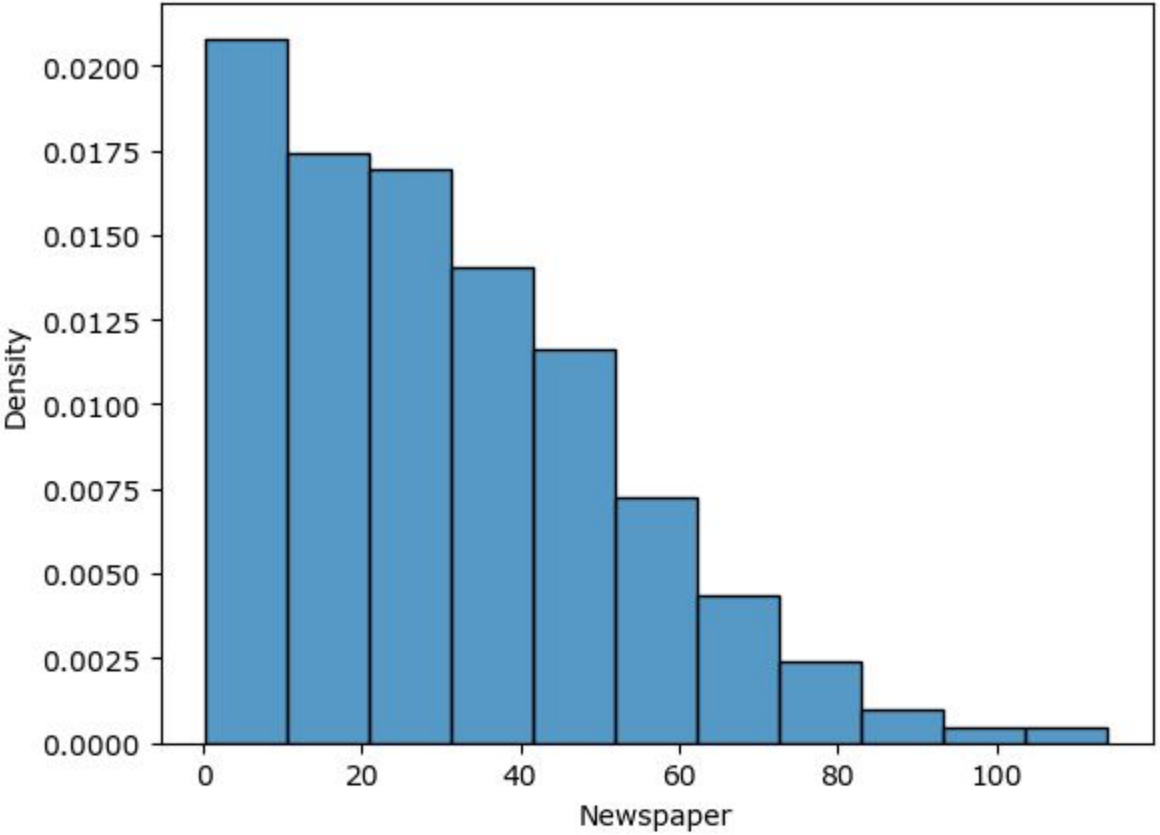
A partir de los datos del ejercicio 1,

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, 23.2, 25.9, 29.7, 35.2 51.2, 54.7, 65.9, 75



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

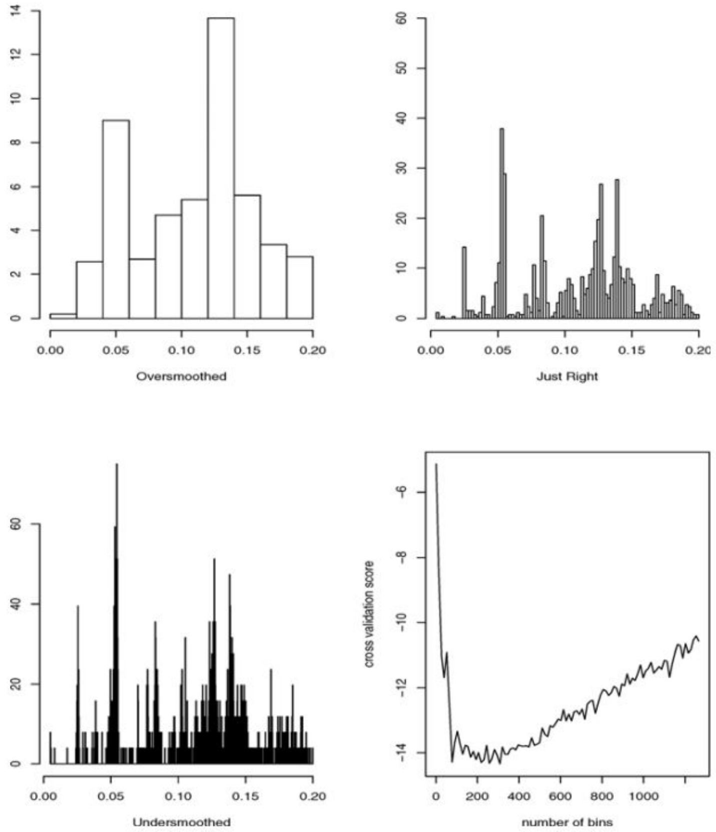


Propiedades del histograma

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

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

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



Estimación de densidad por kernel

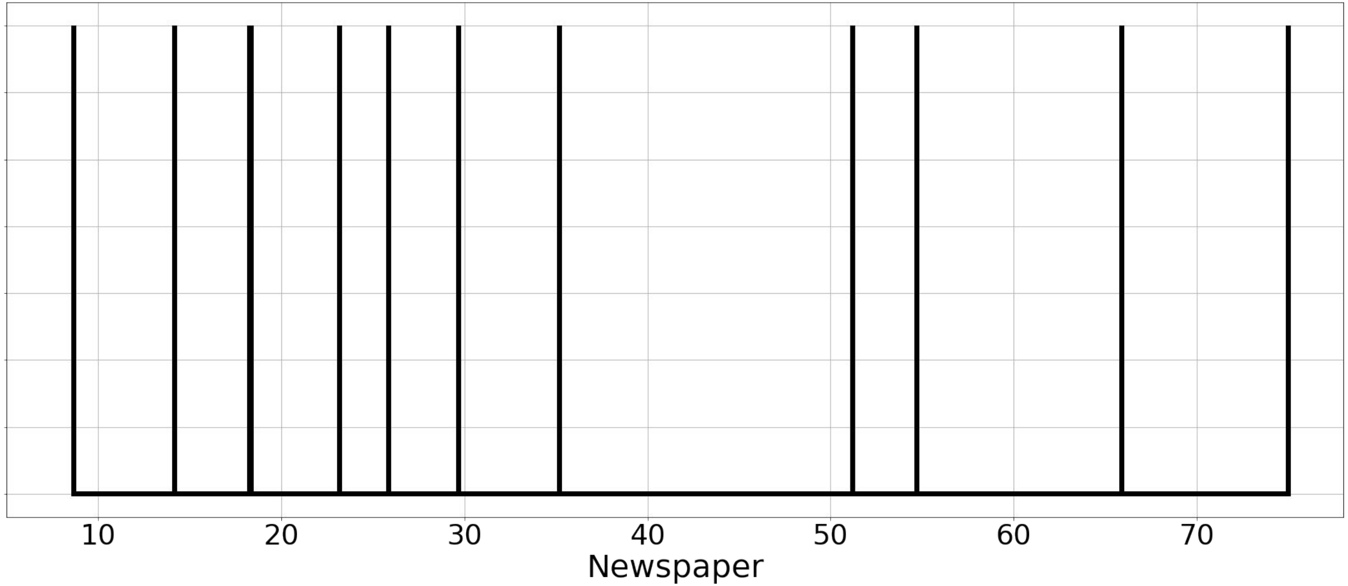
Los histogramas son discontinuos

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

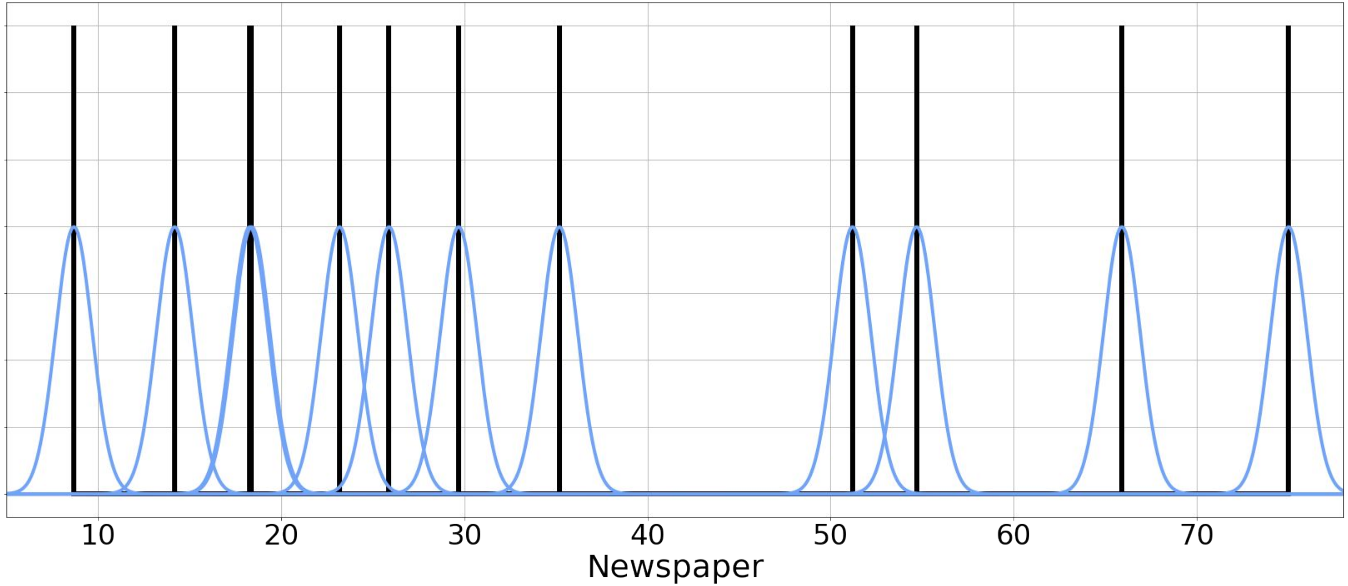
Kernel density estimation

Primero:
marcamos las observaciones en el eje x



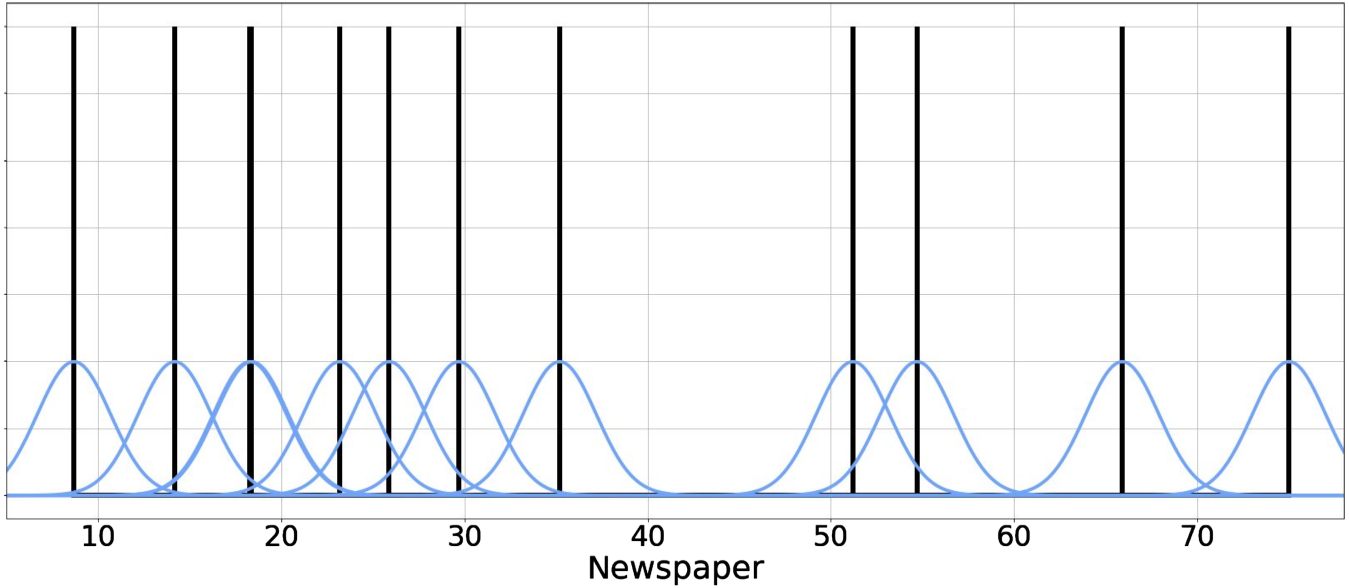
Kernel density estimation

Segundo:
Montamos una función (kernel) sobre cada muestra



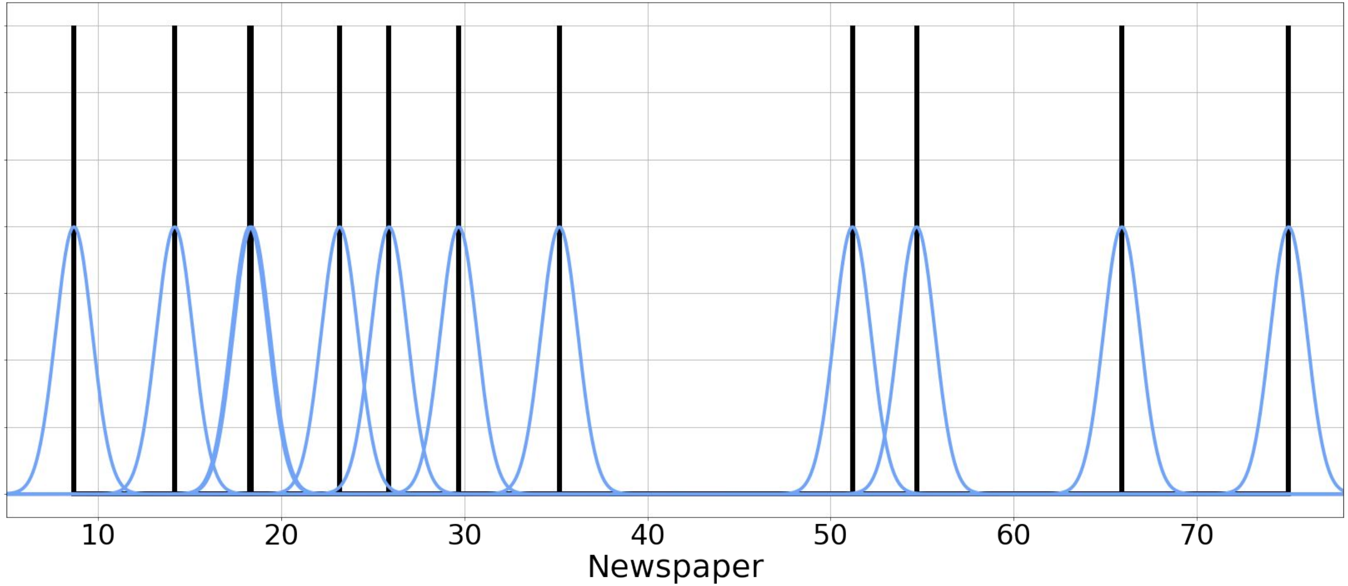
Kernel density estimation

Segundo: Montamos una función (kernel) sobre cada muestra



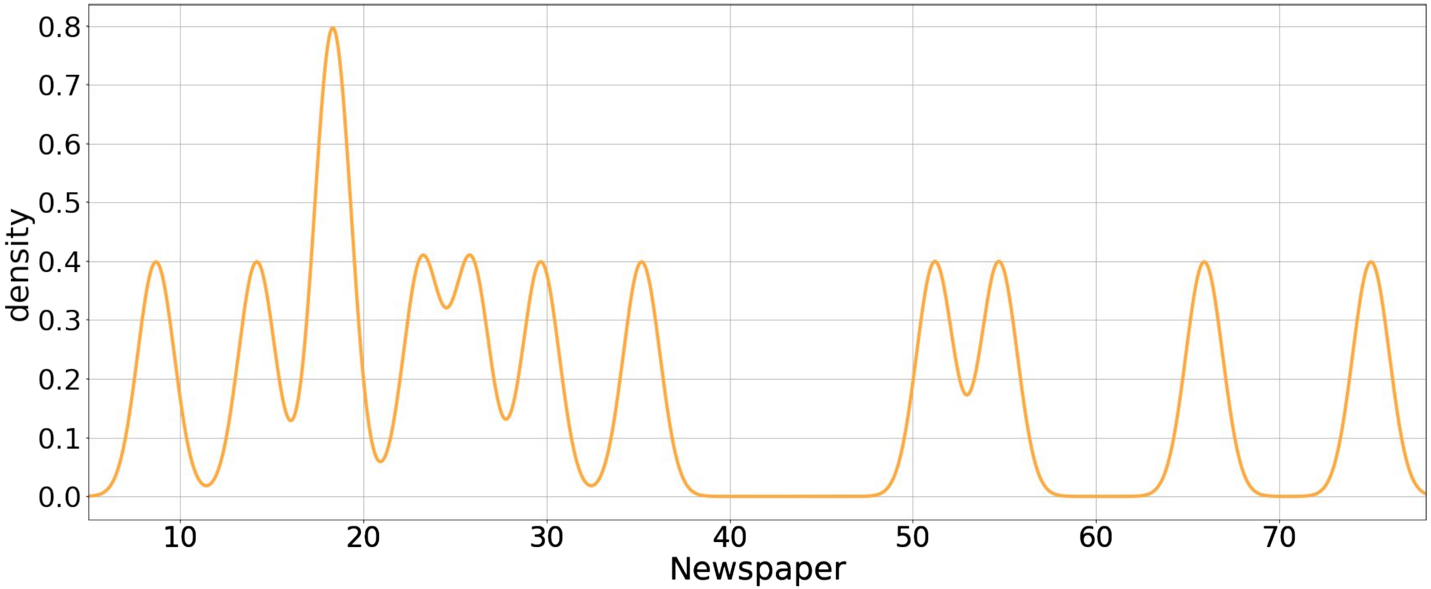
Kernel density estimation

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, 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

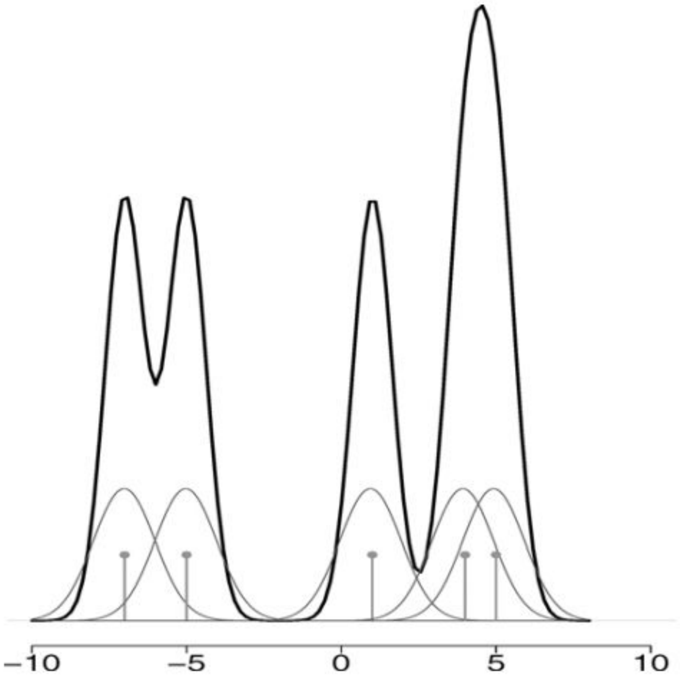
- Gaussiano (simple)

KDE

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

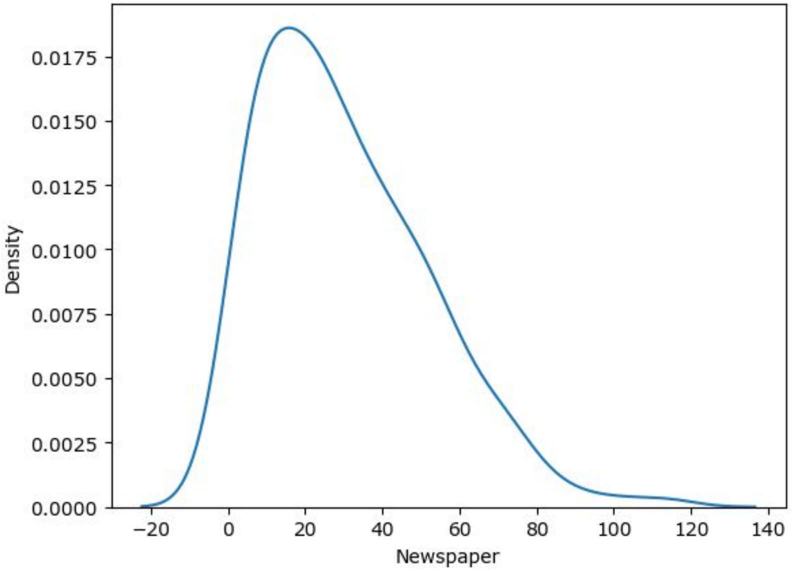
$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}H\left(\frac{x-X_i}{h}\right)$$

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



Ejercicio 3

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



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

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

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 normal de media μ y varianza 4. Hallar una cota inferior del 95% para μ .

Suponer n=20 y μ=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 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: $U(\underline{X}, \Delta) = \frac{\overline{X} - \overline{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0,1)$
- Comparación de medias con varianzas desconocidas e iguales:

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

Bibliografía

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