Probabilidad y estadística

Clase 4 y 5

Estimación Bayesiana

Enfoque Bayesiano

El enfoque Bayesiano supone que se tiene alguna información previa sobre el parámetro, o una ‘creencia’. Esta información previa se expresa en forma de una distribución sobre θ, llamada distribución a priori.

El objetivo del enfoque Bayesiano es ir actualizando nuestra creencia a partir de las muestras observadas.

Distribución a priori

Según el problema se pueden tener distintas interpretaciones acerca de la distribución a priori $\pi(\theta)$:

- La distribución a priori está basada en experiencias previas similares
- La distribución a priori expresa una creencia subjetiva.

A la variable aleatoria la llamaremos Θ , para distinguirla del valor que toma θ.

También cambia la interpretación de familia de distribución, ya que ahora $F_\theta(x)$ es la distribución condicional de X dado que $\Theta = \theta$

Distribución a posteriori

Una vez observada la m.a. $\underline{X} = (X_1, \ldots, X_n)$ se puede calcular la distribución de Θ dada $\underline{X} = \underline{x}$. A esta distribución se la conoce como distribución a posteriori y está dada por

$$f_{\Theta|\underline{X}=\underline{x}}(\theta) = \frac{f_{\underline{X}|\Theta=\theta}(\underline{x})\pi(\theta)}{\int_{-\infty}^{\infty} f_{\underline{X}|\Theta=\theta}(\underline{x})\pi(\theta)d\theta}.$$

Obs: Si alguna de las variables ($\underline{X}$ o Θ) son discretas reemplazaremos la función de densidad por una función de probabilidad, y si Θ reemplazaremos la integral por una sumatoria.

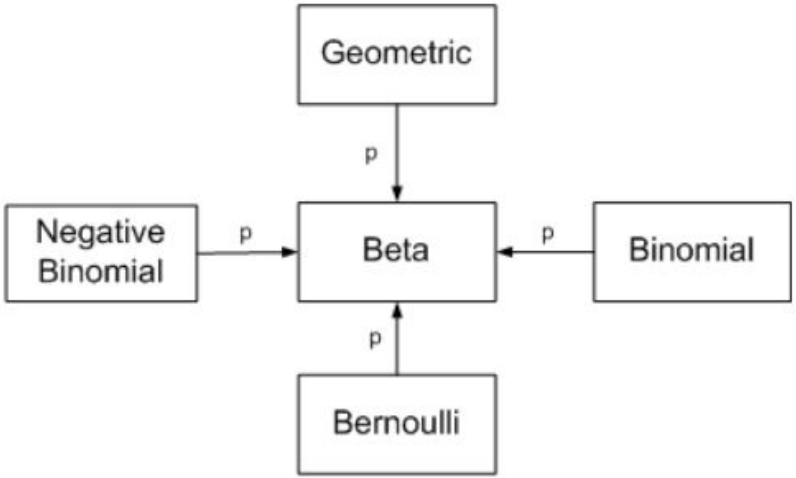
Cómo elegir la distribución a priori

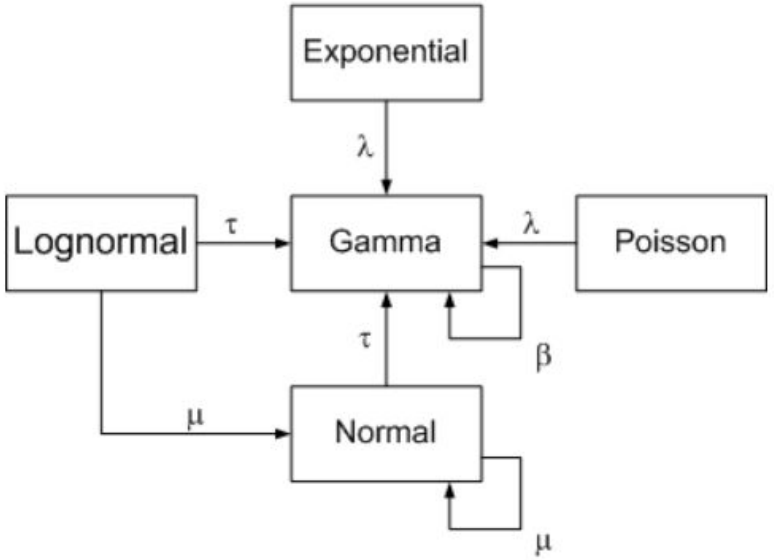
- Cuando no tenemos información acerca del parámetro vamos a usar distribuciones a priori que no favorezcan ningún valor, como por ejemplo una distribución uniforme.

- Hay que prestar atención al soporte que tiene la distribución a priori elegida, ya que los valores de θ que no pertenezcan al soporte original quedan "eliminados".

- Por simplicidad, suelen elegirse **distribuciones conjugadas**. Las familias de distribuciones conjugadas son aquellas para las cuales la función de dist. a posteriori va a pertenecer a la misma familia de distribuciones que la dist. a priori.

Ejemplos de familias de dist. conjugadas





Fuente

Ejercicio 1

La posición del impacto en un tiro al blanco (en decímetros) respecto del cero sobre el eje x es una variable aleatoria X con distribución normal de media cero y varianza 1/θ, donde θ representa la precisión del tirador.
A priori, la precisión θ tiene una distribución Chi-cuadrado de 8 grados de libertad. Lucas tiro 10 veces al blanco y observó que $\sum_{i=1}^{10} x_i^2 = 17$. Hallar la distribución a posteriori de θ.

Ejercicio 2

La cantidad de accidentes semanales en una planta industrial tiene una distribución de Poisson de media μ.

En una muestra de 100 semanas se observaron las frecuencias:

Cantidad de accidentes	0	1	2	3	4	5
Frecuencia	10	29	25	17	13	6

(hubo 10 semanas sin accidentes, hubo 29 semanas con 1 accidente...)

A priori, μ tiene una distribución exponencial de media 2. Hallar la distribución a posteriori de μ.

Estimadores puntuales Bayesianos

Una ventaja de este método es que se pueden definir de manera natural estimadores óptimos. Según que se considere como función de pérdida se obtendrán distintos estimadores puntuales.

Si consideramos una función de pérdidas $l(\theta, d)$, que es el costo de estimar al parámetro θ por el valor d, si $\hat{\theta} = \varphi(\underline{X})$, entonces la pérdida será una v.a. $l(\Theta, \varphi(\underline{X}))$.

La pérdida esperada es lo que se conoce como riesgo de Bayes. Esto significa que dada una distribución a priori $\pi(\theta)$, un estimador de Bayes será el que minimice $r(\varphi, \pi) = \mathbb{E}[l(\Theta, \varphi(\underline{X}))]$

Distintos estimadores según la función de riesgo

- Si consideramos una función de pérdida cuadrática: $\ell(\theta, d) = (\theta - d)^2$, el estimador de Bayes será es que minimice el ECM. ¿Quién era este estimador? $\varphi(\underline{X}) = \mathbb{E}[\Theta|\underline{X}]$ (con la esperanza tomada respecto de la distribución a posteriori.)

- Si consideramos la pérdida ℓ_1 dada por $\ell(\theta, d) = |\theta - d|$, el estimador de Bayes será la mediana de la distribución a posteriori de Θ condicionada a $\underline{X} = \underline{x}$.

- Muchas veces se utiliza lo que se conoce como máximo a posteriori (MAP) que se corresponde con la moda de la distribución a posteriori de $\Theta|\underline{X} = \underline{x}$, es decir el de el valor de θ que maximiza la dist. a posteriori. En general no es un estimador Bayesiano.

Ejercicio 3

Para el ejercicio 1, hallar la estimación de Bayes de θ para el riesgo cuadrático.

Cálculo de probabilidades

Para poder estimar probabilidades a partir de la distribución a posteriori de los parámetros aleatorios, usaremos la fórmula de probabilidad total. Por ejemplo, si Θ y X son v.a. continuas

$$\mathbb{P}(X \in A) = \int_A \int_{-\infty}^{\infty} f_{X|\Theta=\theta}(x) f_{\Theta|\underline{X}=\underline{x}}(\theta) d\theta dx$$

Ejercicio 4

Para el ejercicio 2, estimar la probabilidad de que en la semana del 18 de diciembre de 2021 no ocurra ningún accidente en la mencionada planta.

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 $x \in \mathbb{R}$,

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

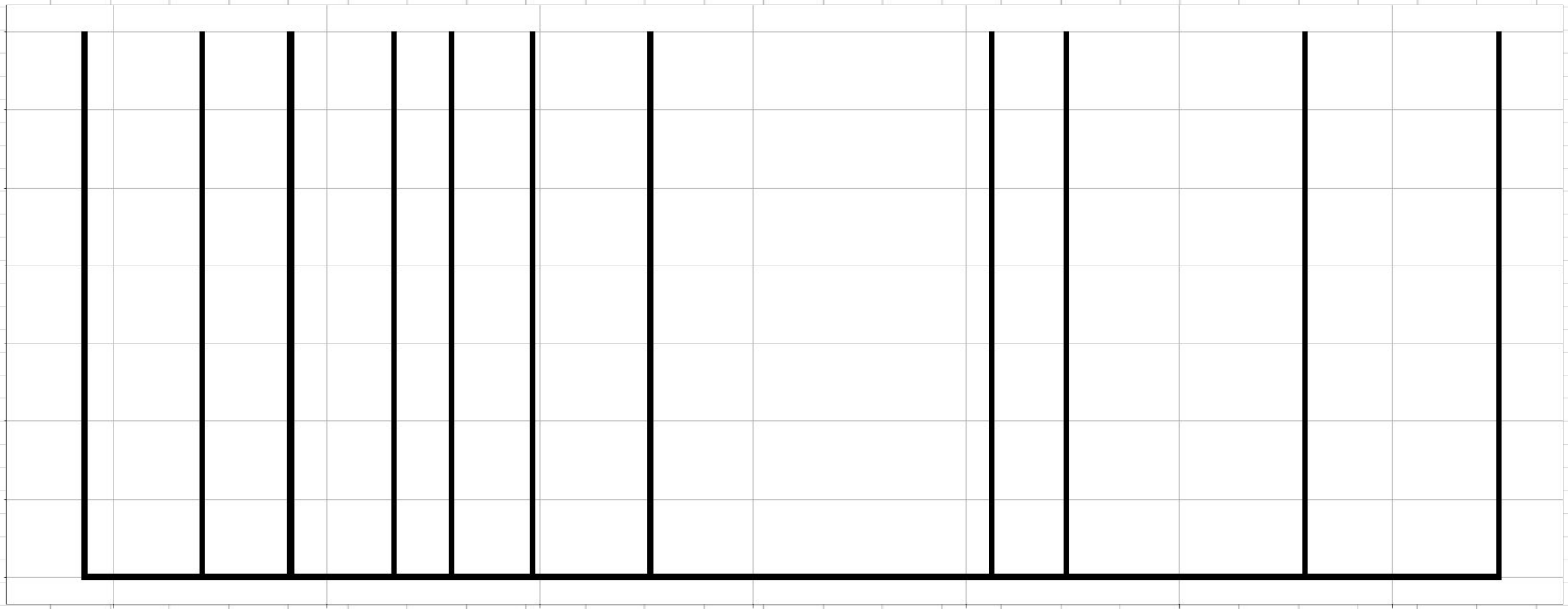
All of Statistics, Wasserman

Ejercicio 1

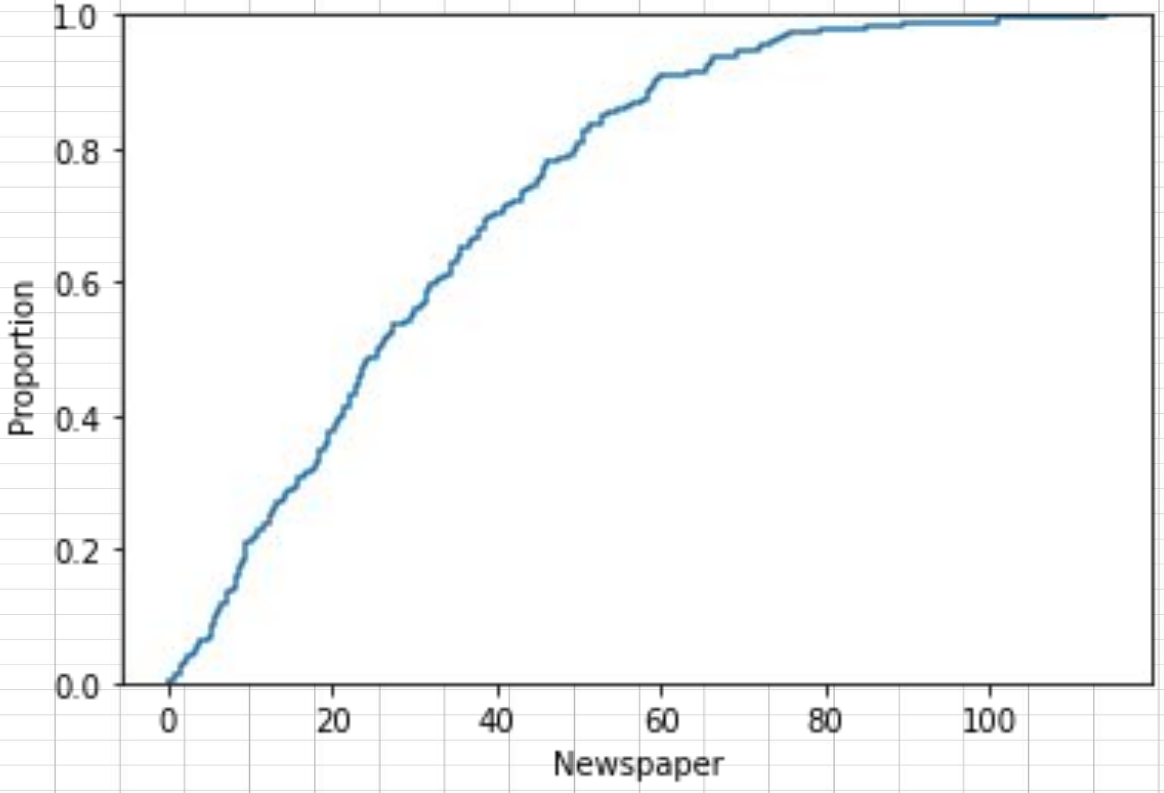
Usemos el Advertising Sales Dataset. Allí se presentan valores del presupuesto asignado (en 1000$) en distintos medios (TV, radio, diarios) y las ventas asociadas.

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)



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)

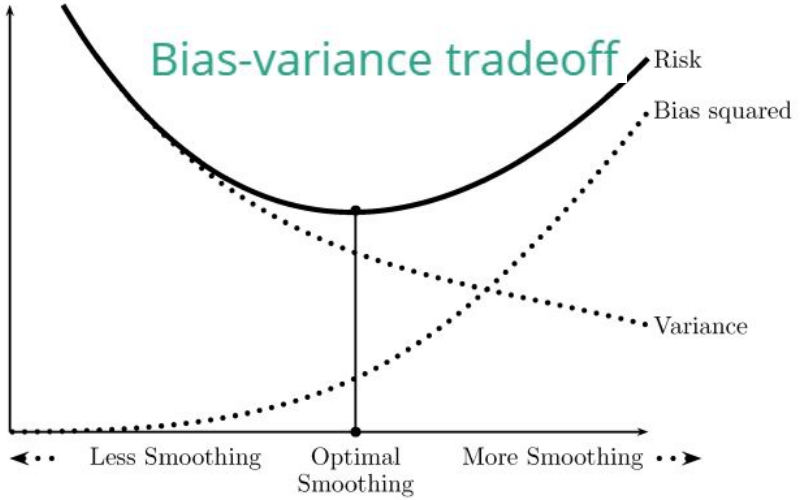
Para densidades vamos a definir el riesgo:

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

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

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

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



All of Statistics, Wasserman

Histogramas

1. Se toman los valores máximo y mínimo y se divide el intervalo en m subintervalos 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.

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

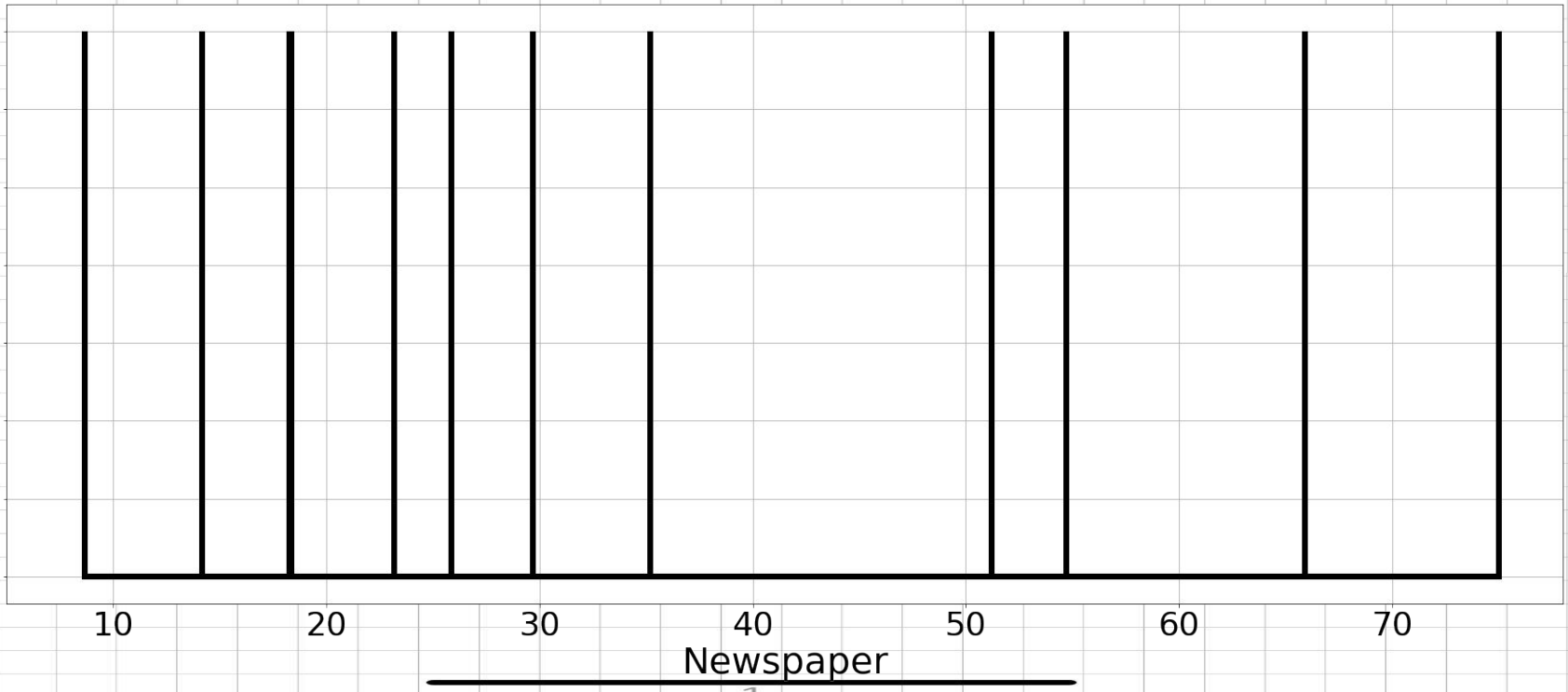
$$\hat{f}_n(x) = \frac{1}{h} \sum_{j=1}^{m} \hat{p}_j 1\{x \in B_j\} \quad \text{donde} \quad \hat{p}_j = \nu_j / n$$

Ejercicio 2

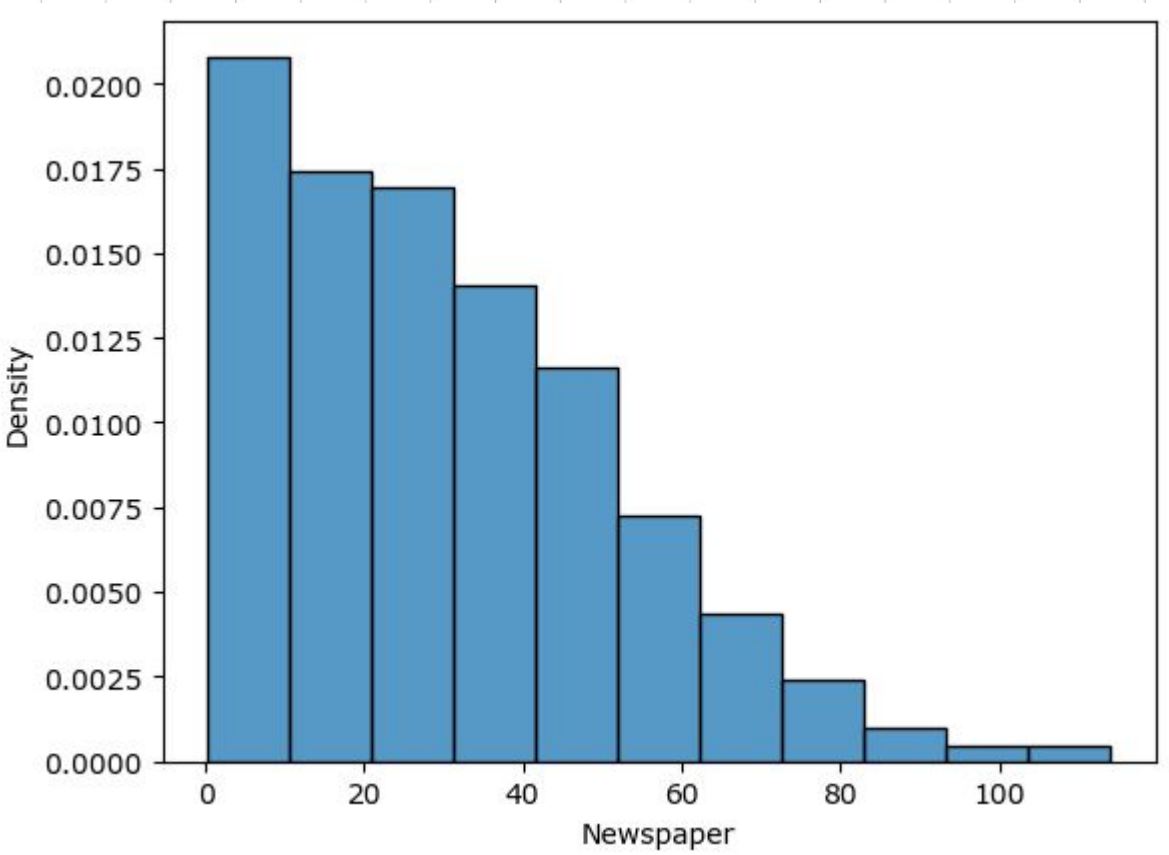
A partir de los datos del ejercicio 1,

1. Calcular a mano, el histograma de 6 bins (subintervalos)
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



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

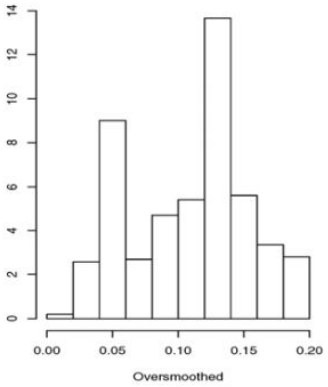


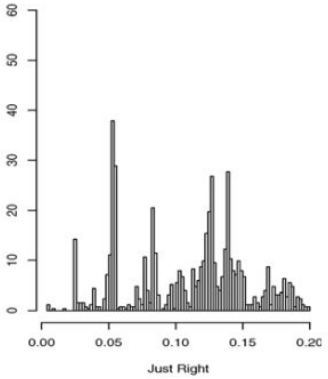
Propiedades del histograma

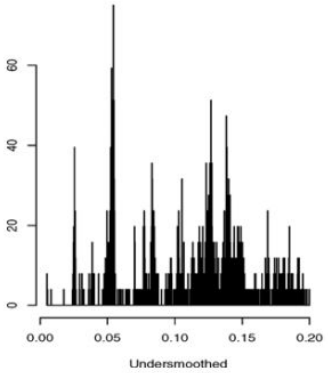
Teorema: Sea x y m fijos, y sea B_n el bin que contiene a x, luego

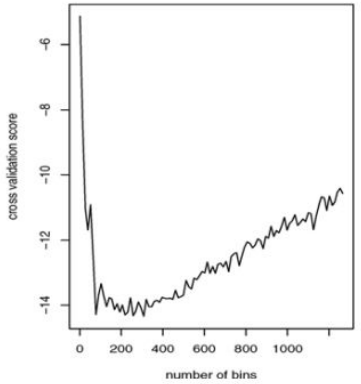
$$\mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h} \qquad \mathbb{V}(\widehat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}.$$

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









Estimación de densidad por kernel

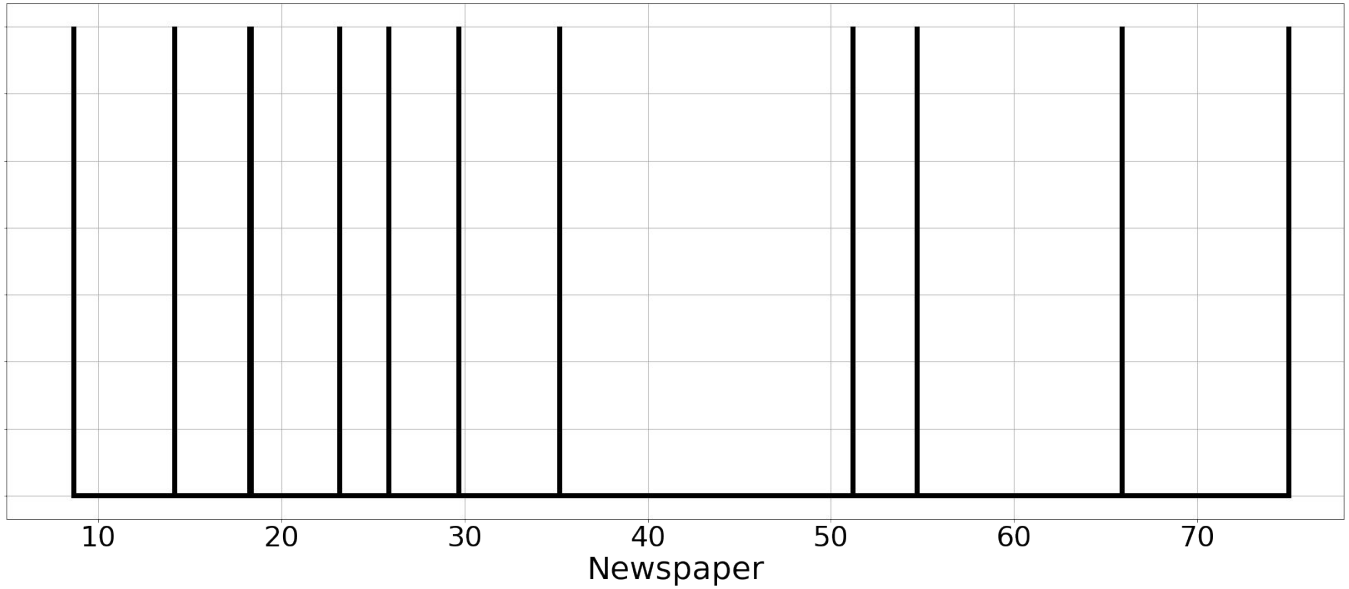
Los histogramas son discontinuos

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

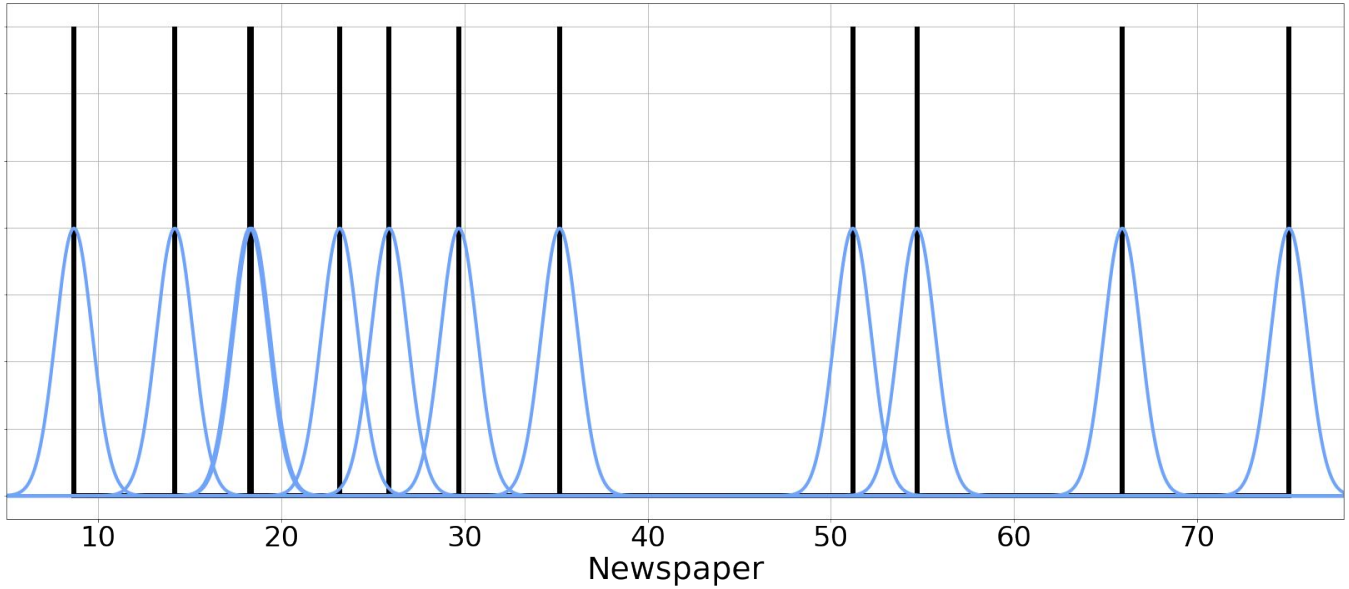
Kernel density estimation

Primero:
marcamos las observaciones en el eje x



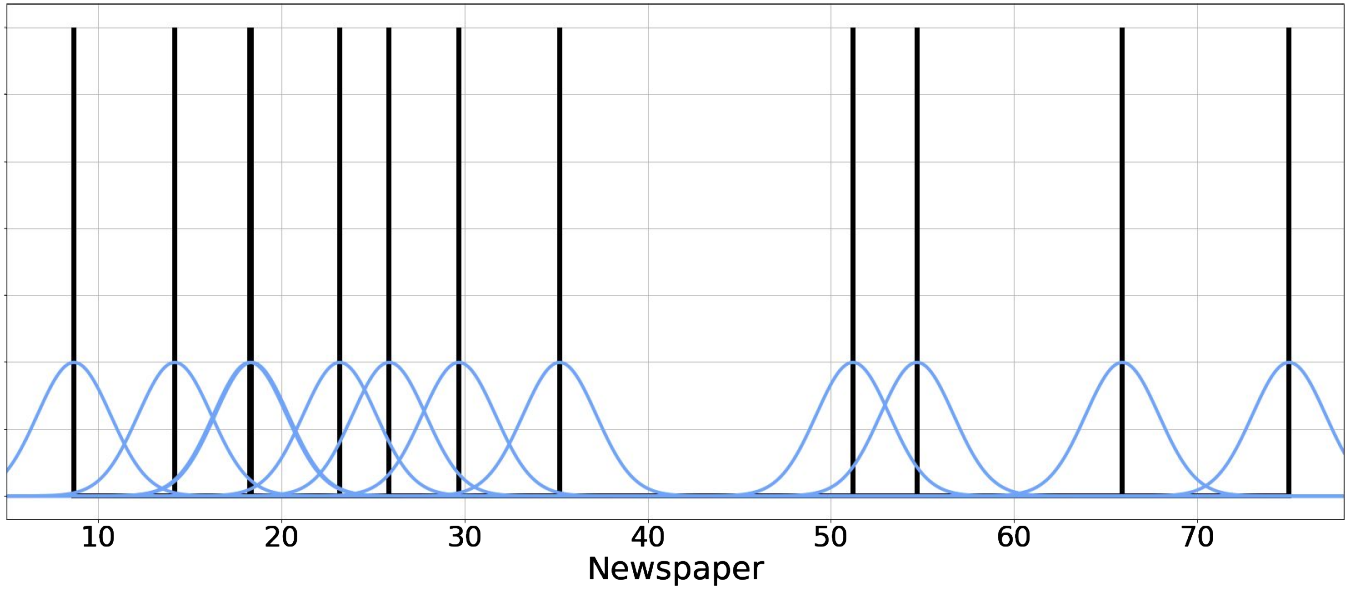
Kernel density estimation

Segundo:
Montamos una función (kernel) sobre cada muestra



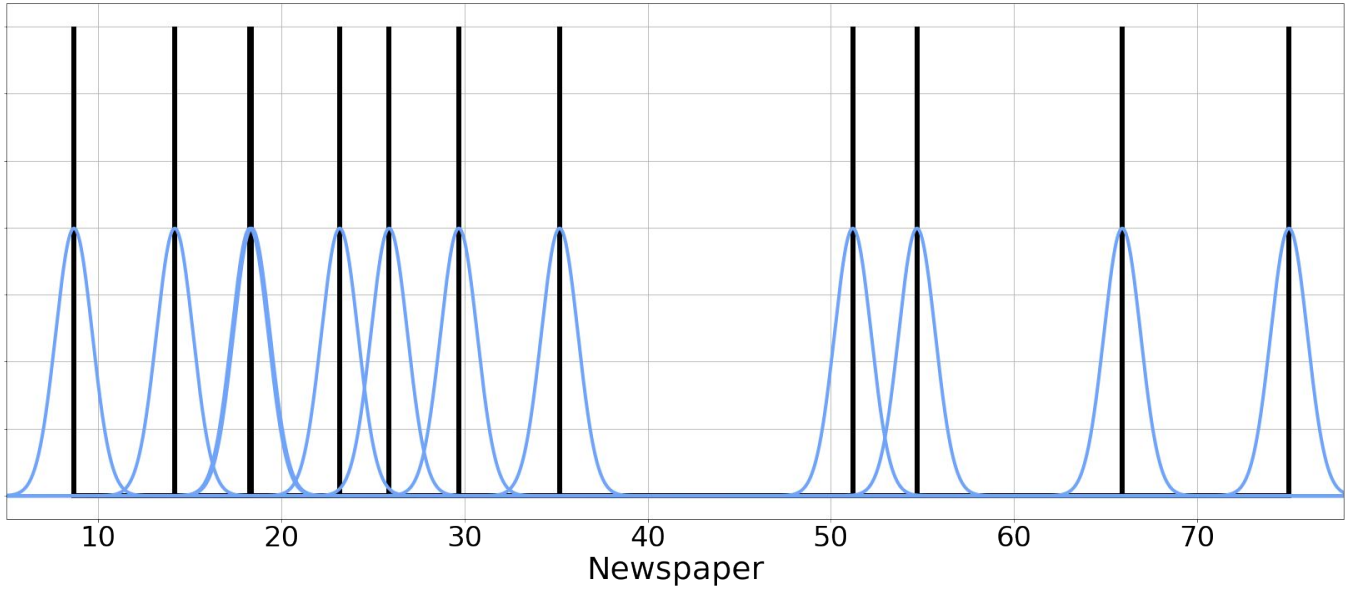
Kernel density estimation

Segundo:
Montamos una función (kernel) sobre cada muestra



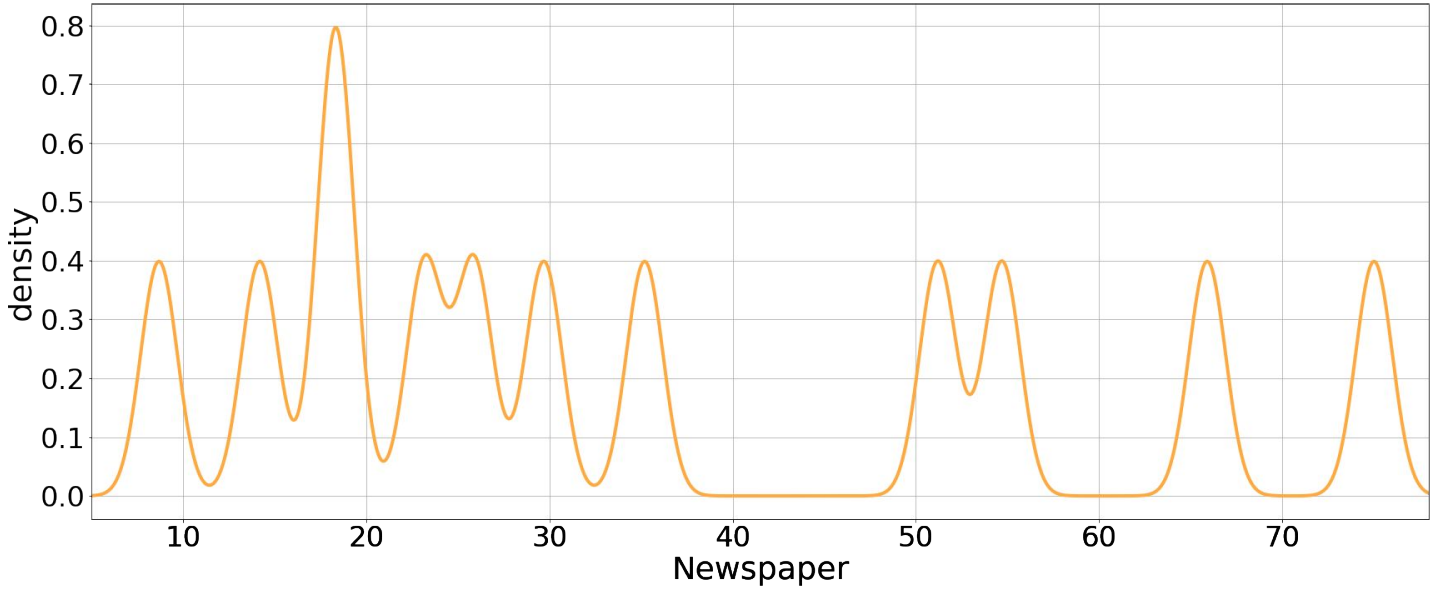
Kernel density estimation

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

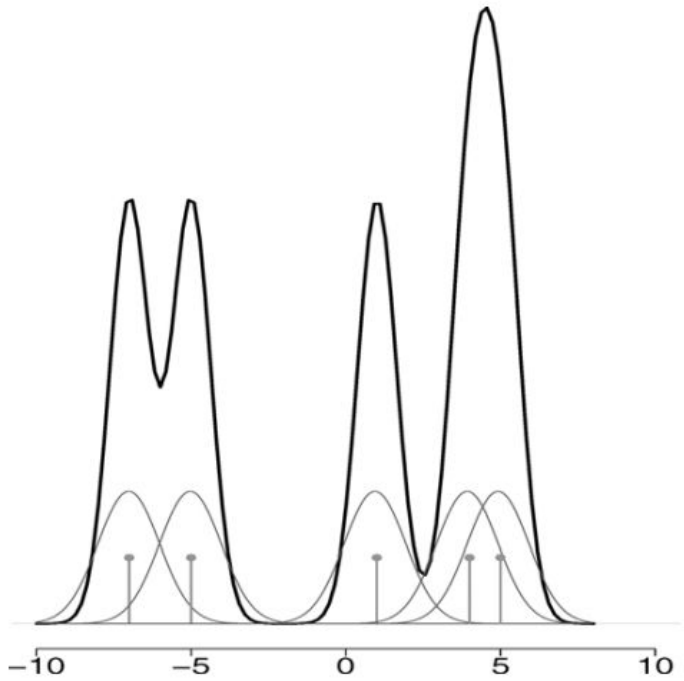
- Gaussiano (simple)

KDE

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

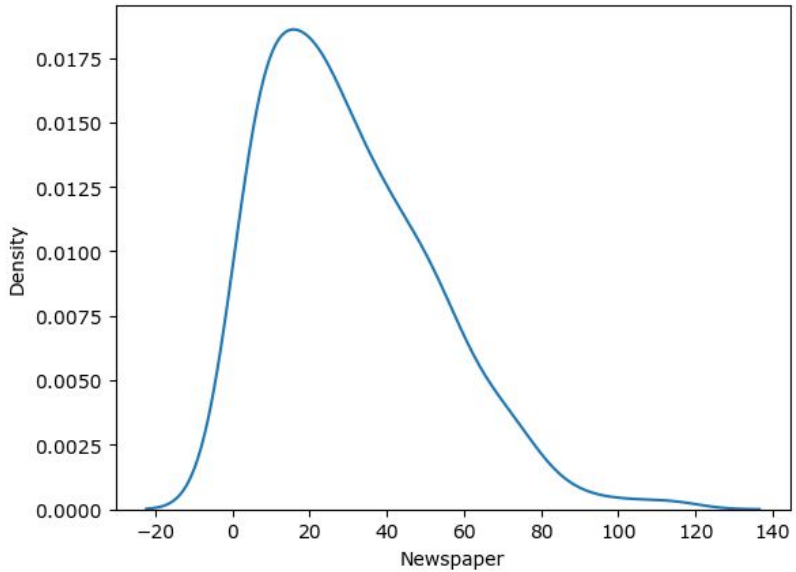
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} H\left(\frac{x - X_i}{h}\right)$$

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.



Regresión no paramétrica

$$\hat{f}\,(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

$$\hat{f}\,(x, y) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) K_h(y - y_i)$$

Regresión no paramétrica

$$E(Y \mid X = x) = \int y f(y \mid x)\, dy = \int y \frac{f(x,y)}{f(x)}\, dy$$

Estimador de Nadaraya–Watson [1964]:

$$\hat{E}(Y \mid X = x) = \int \frac{y \sum_{i=1}^{n} K_h(x - x_i) K_h(y - y_i)}{\sum_{j=1}^{n} K_h(x - x_j)}\, dy,$$

$$= \frac{\sum_{i=1}^{n} K_h(x - x_i) \int y\, K_h(y - y_i)\, dy}{\sum_{j=1}^{n} K_h(x - x_j)},$$

$$= \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{j=1}^{n} K_h(x - x_j)},$$

Regresión no paramétrica

Tarea: Usar las columnas ‘TV’, ‘Radio’ y ‘Newspaper’ por separado para predecir ‘Sales’ usando los datos ‘advertising.csv’ con el estimador de Nadaraya–Watson. (Usar como ancho de banda la mediana de la distancia entre pares de muestras).

1) estimar el error cuadrático medio
2) ¿cuál es el medio de comunicación donde las inversiones logran predecir mejor las ventas?

Bibliografía

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