# Probabilidad y estadística

## Clase 5

# Estimación no paramétrica

# Función de distribución empírica

$F(x) = P(X \leq x)$ $\forall x$

Tenemos tal que

$P(x)$ empírica

$\frac{1}{n}$ $\frac{1}{n}$ ....

$x_1$ $x_2$

$X_1, \ldots, X_n$ una m.a.

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:

es una uniforme con soporte en $\{x_i; i=1,\ldots,n\}$

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

# Propiedades de la ECDF

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

$$
\begin{aligned}
\mathbb{E}\left(\widehat{F}_n(x)\right) &= F(x), \\
\mathbb{V}\left(\widehat{F}_n(x)\right) &= \frac{F(x)(1-F(x))}{n}, \qquad O\left(\tfrac{1}{n}\right) \\
\text{Bias}^2 + \text{Var} \quad = \text{MSE} &= \frac{F(x)(1-F(x))}{n} \to 0, \\
\widehat{F}_n(x) &\xrightarrow{\text{P}} F(x).
\end{aligned}
$$

$$\mathbb{P}\left(\left|\hat{F}_n(x) - F(x)\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \quad \forall\, \varepsilon > 0$$

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.

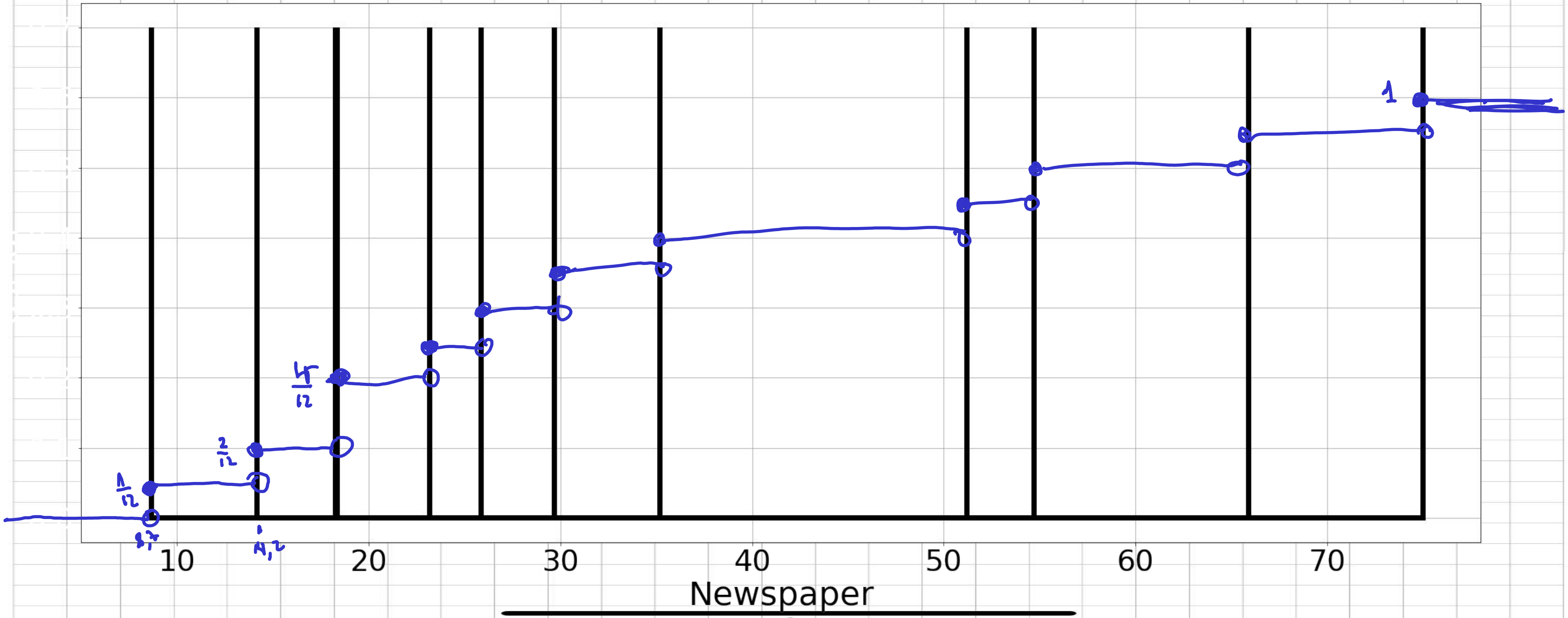
1. A partir de la muestra 8.7, 14.2, 18.3, 18.4, 23.2, 25.9, 29.7, 35.2 51.2, 54.7, 65.9, 75 obtener la función de distribución empírica a mano.

   } cuánto se gasta en publicidad TV.

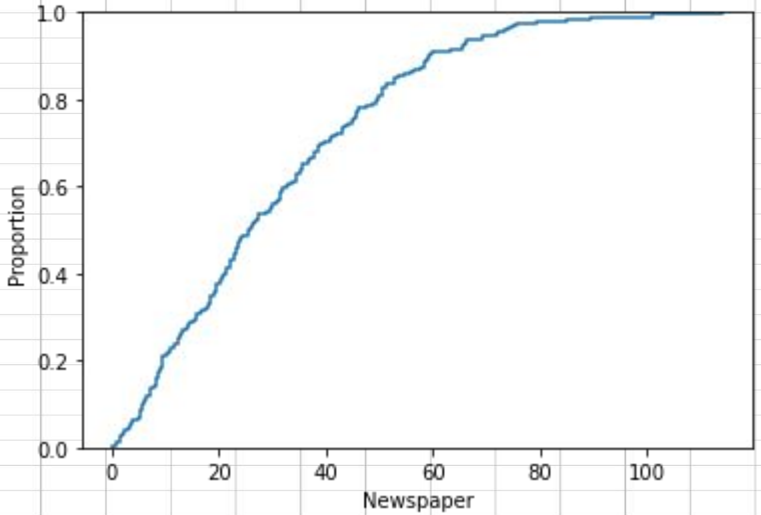
2. Utilizar la columna “Newspaper” del archivo “advertising.csv” y calcular la func. de distribución empírica usando Python.

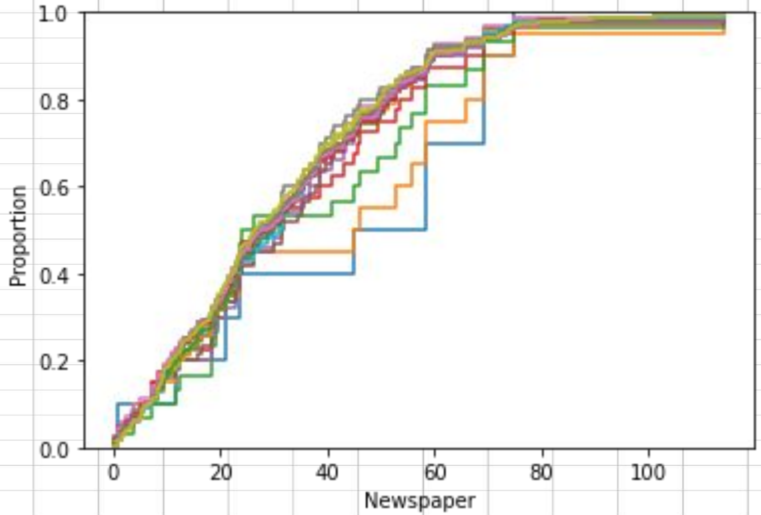
A partir de la muestra 8.7, 14.2, 18.3, 18.4, 23.2, 25.9, 29.7, 35.2 51.2, 54.7, 65.9, 75 obtener la función de distribución empírica a mano.

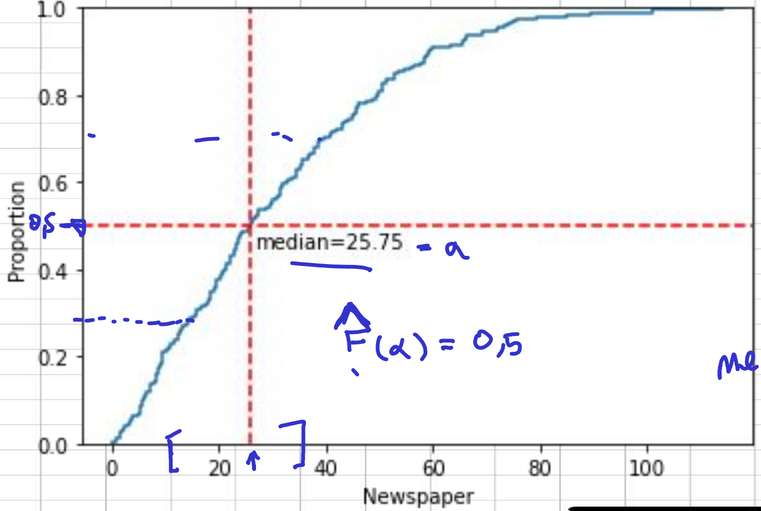
$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\} = 0$$



sns.ecdfplot(newspaper)




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

$\sum_{x \in \{x_1, \ldots, x_n\}} x \frac{1}{n}$

obs:

mediana $= \hat{F}_{n(0,5)}^{-1} = \inf\{x : \hat{F}_n(x) \geq p\}$

# Estimación de densidades (*smoothing*)

$\text{bias}^2$ ECM Var

$$\underline{ECM} = E[(\hat{\theta} - \theta)^2] = \text{Bias}^2 + \text{Var} \rightarrow$$
$$= \langle \hat{\theta} - \theta, \hat{\theta} - \theta \rangle = \|\theta - \hat{\theta}\|^2$$

$\boxed{\text{Espacio de Hilbert}}$ p.i.
$\langle a,b \rangle = E[ab]$
$\text{Var}(X) < \infty$
$\int (x-\mu)^2 f(x)\,dx$

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

{E.H. de la funciones de cuadrado integrables" $\Leftrightarrow \{f : \int f^2(x)dx < \infty\}$ y definimos el p.i. $\langle f,g \rangle = E\left[\int f \cdot g \, dx\right]$

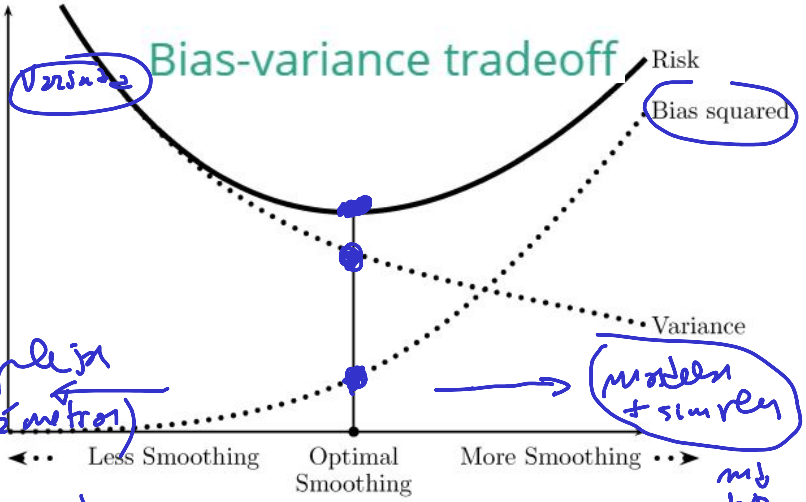
Para densidades vamos a definir el riesgo: = ECMI

$\{x_1, \ldots, x_n\}$ una m.a.

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

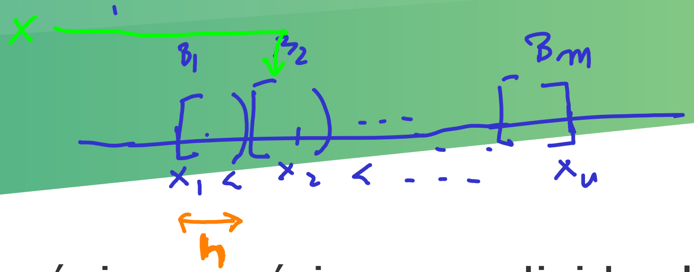
$$= \int_{-\infty}^{\infty} b^2(x)dx + \int_{-\infty}^{\infty} v(x)dx$$

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




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 $h$.

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

$$\int_{-\infty}^{\infty} \hat{f}_n(x)\, dx = 1$$

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

Si definimos $p_J = P(X \in B_J) = \int_{B_J} f(x)\, dx$

$\forall x$, $E[\hat{f}_n(x)] = \frac{E\, \hat{P}_{J_0}}{h} = \frac{P_{J_0}}{h} = \frac{\int_{B_{J_0}} f(x)\, dx}{h} \approx \frac{f(x) \cdot h}{h} = f(x)$

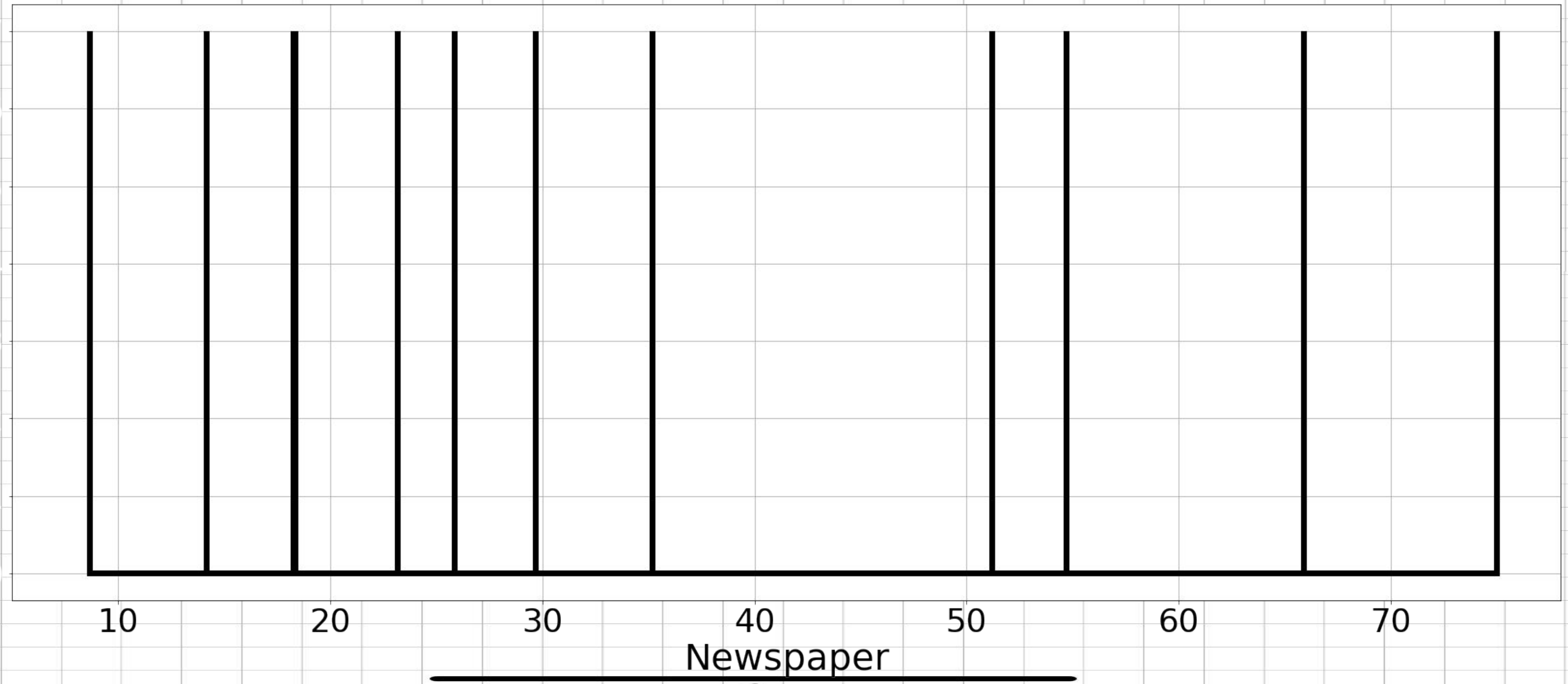
m

# Ejercicio 2

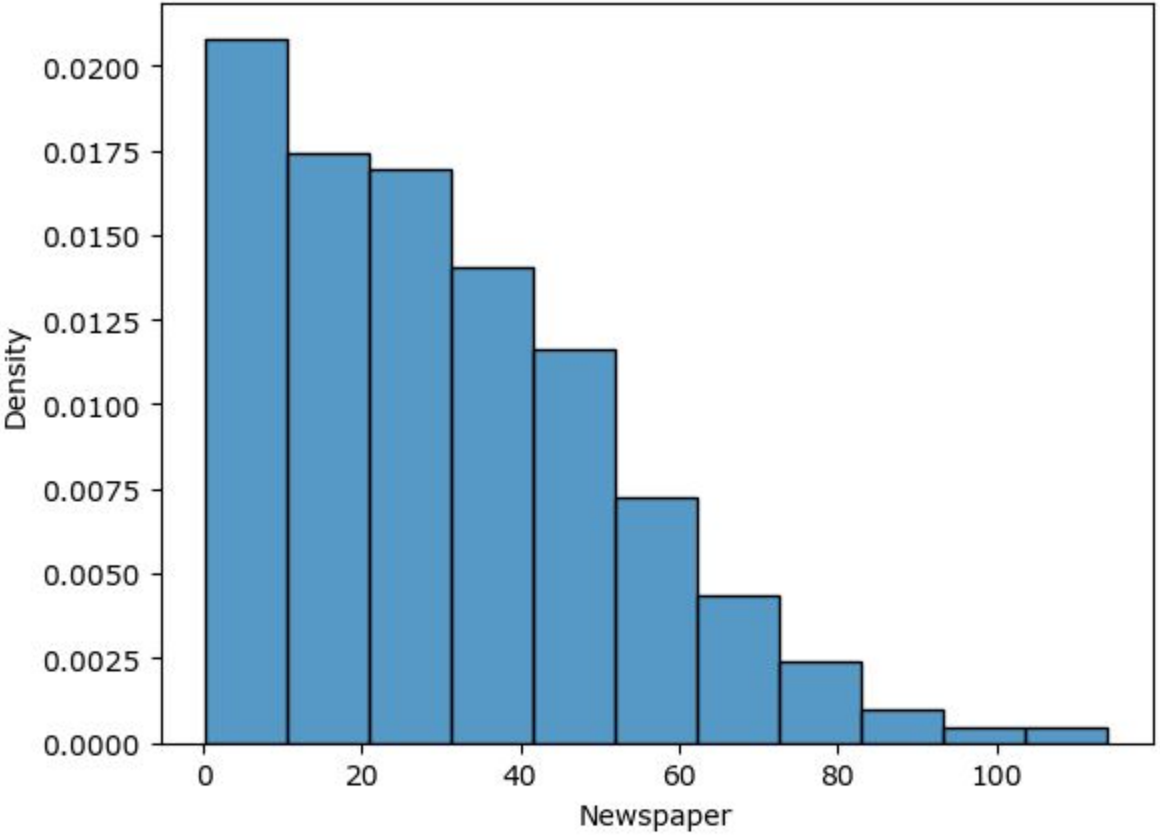
A partir de los datos del ejercicio 1,

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



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



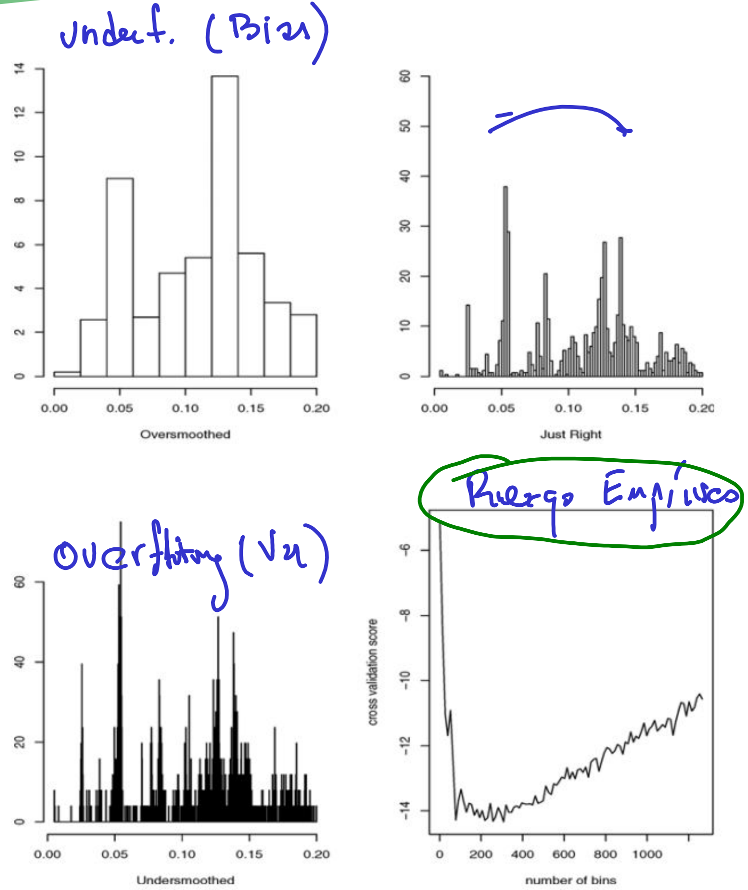
# Propiedades del histograma

nº => sesgo

**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

X es de varianza finita
⇓
suavidad de $f_X(x)$.

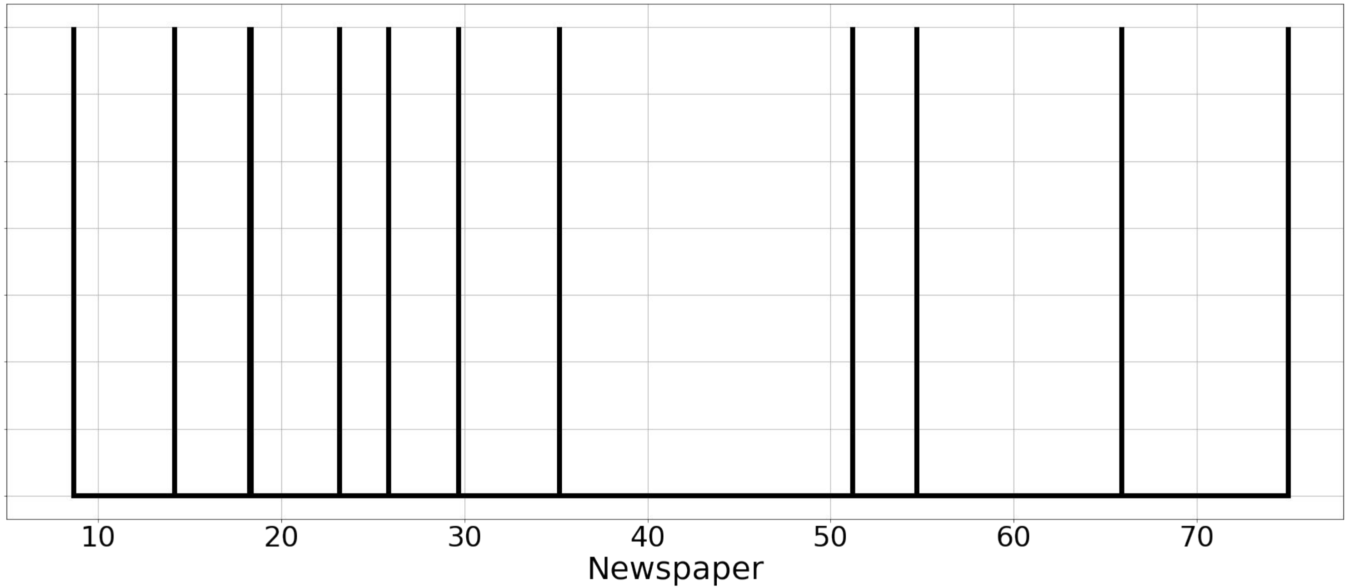
Los histogramas son discontinuos

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

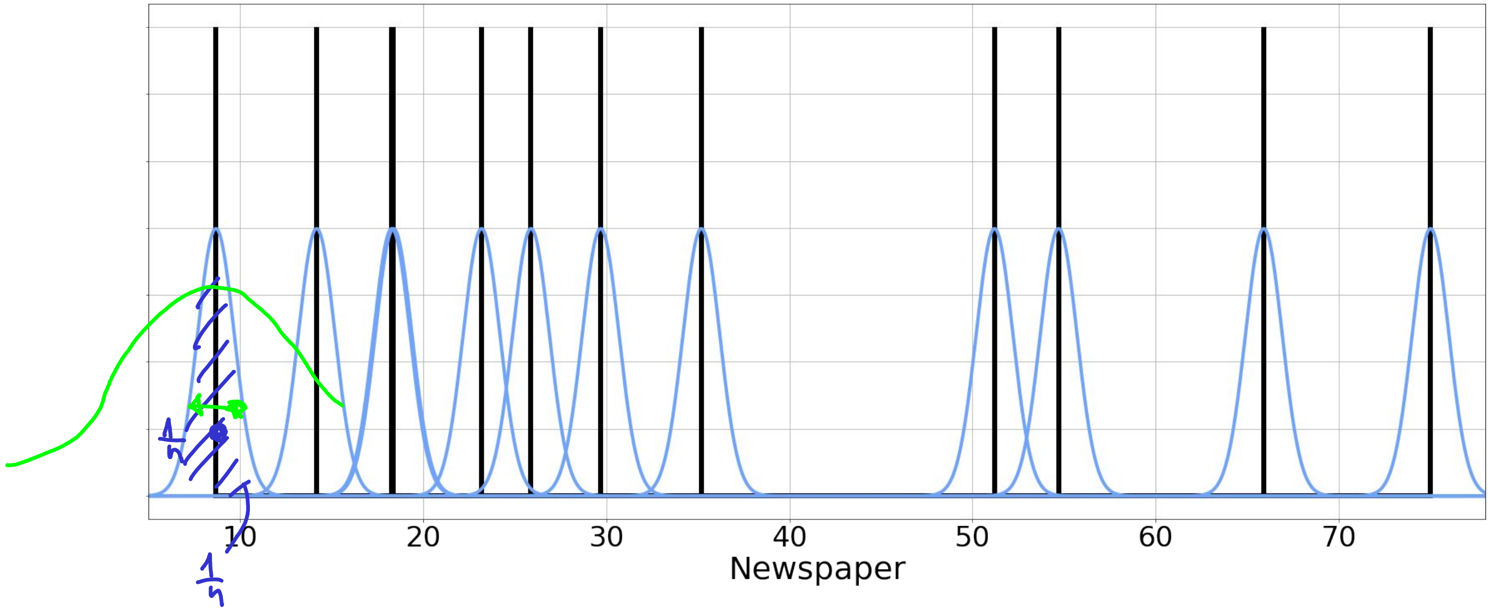
# Kernel density estimation

**Primero:**
marcamos las observaciones en el eje x



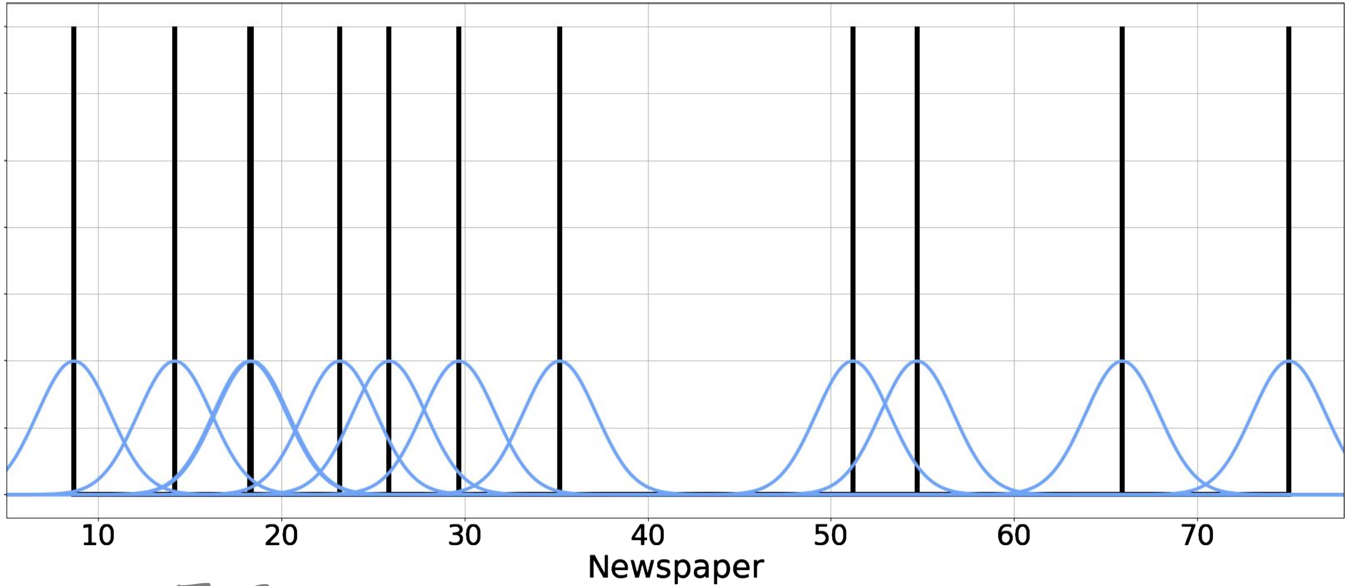
# Kernel density estimation

**Segundo**: Montamos una función (kernel) sobre cada muestra



# Kernel density estimation

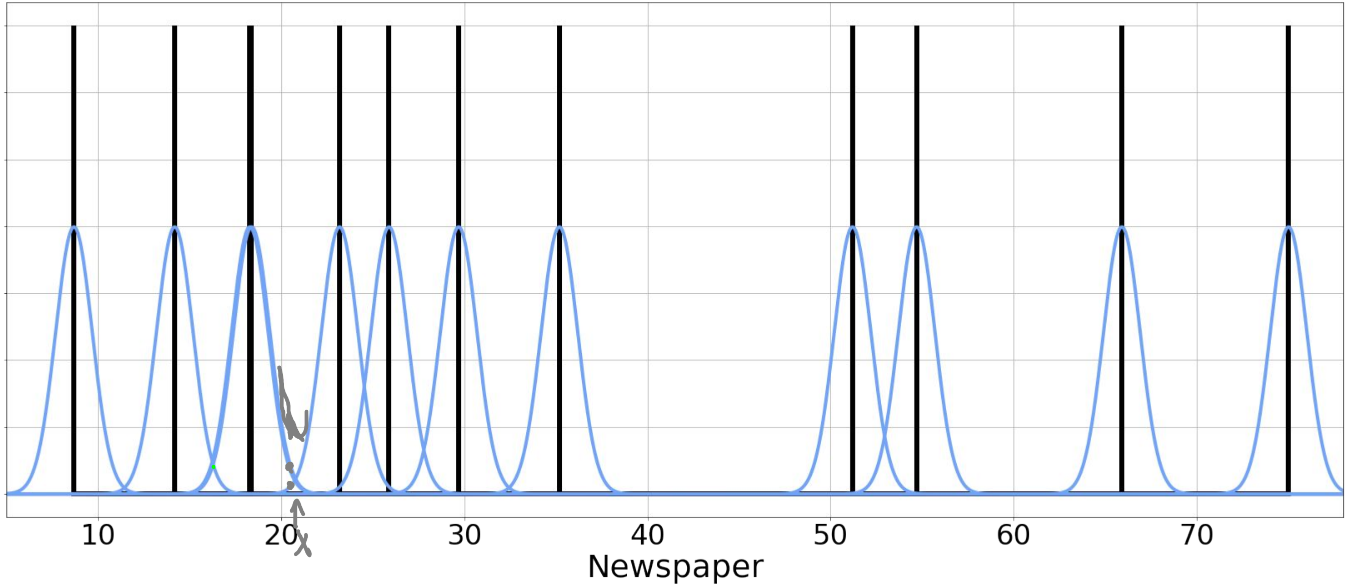
**Segundo**: Montamos una función (kernel) sobre cada muestra




$\} \; K(u) \circledast \sum \frac{1}{n} \delta(x - x_i)$

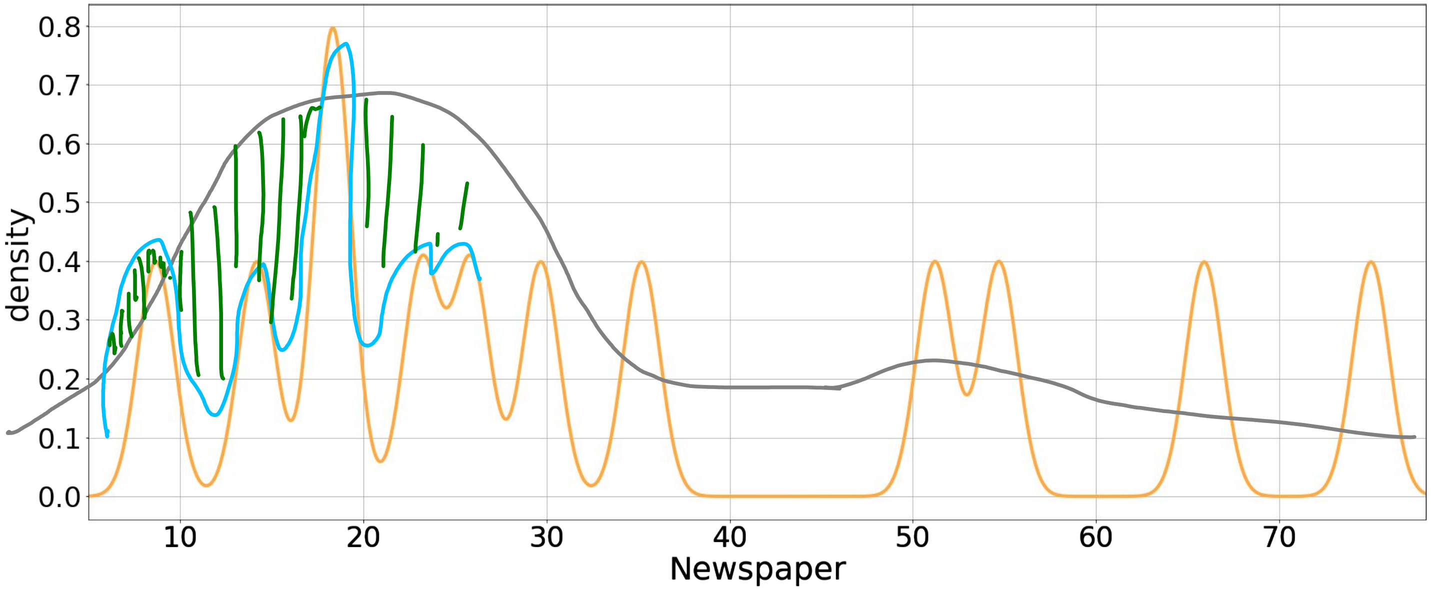
# Kernel density estimation

**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:

$K(x) \geq 0$, $\int K(x)dx = 1$, $\int x \underbrace{K(x)}_{f_X(x)} dx$=0, y $= E_K(x) = 0$

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

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

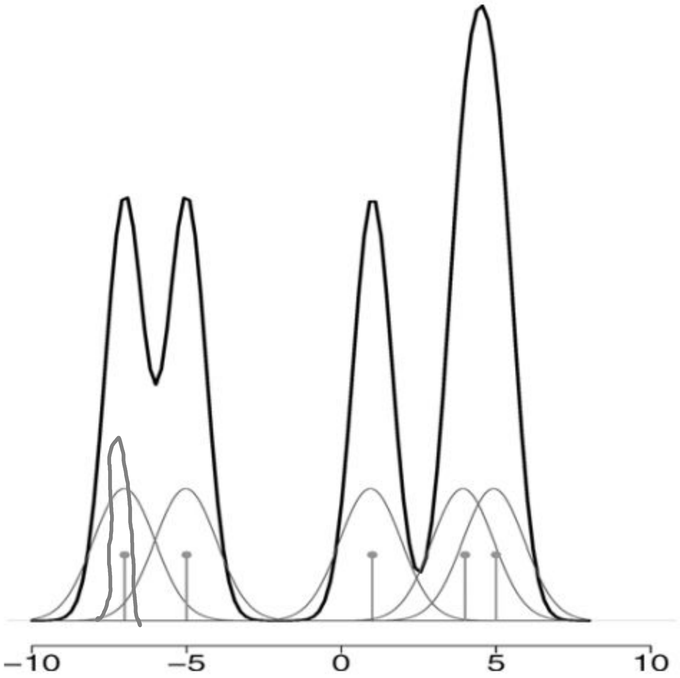
- Gaussiano (simple)

# KDE

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

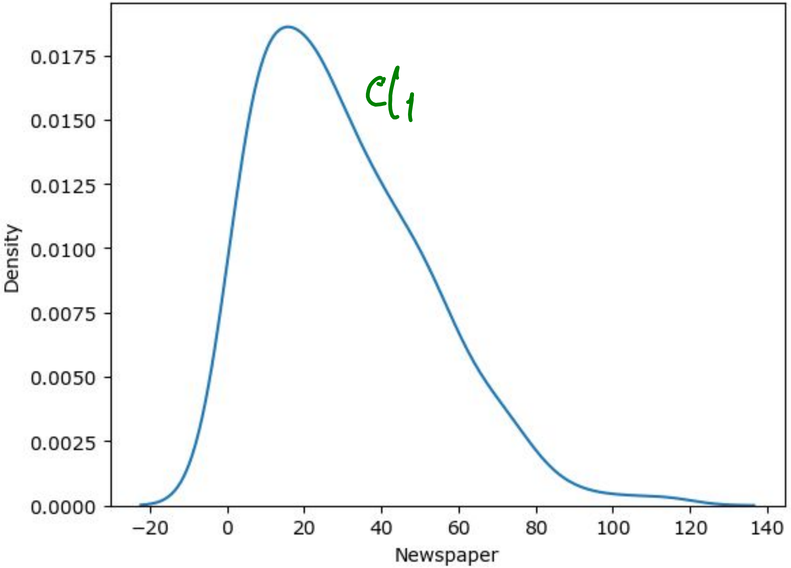
$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{\frac{1}{h}\underset{\overset{\uparrow}{h}}{H}\left(\frac{x-X_i}{h}\right)}^{K(x)}$$

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.




$$\hat{Var}(X) = \int (x - Ex)^2 \hat{f}_n(x)\,dx$$

# Intervalos de confianza

# Motivación

$F_\theta(x)$ quiero estimar $\theta$ basándome en una m.a. $\{x_1, \ldots, x_n\}$.

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 $\theta$.

$P(\theta \in [a,b]) \approx 1$

$[a_n, b_n]$

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

# Región de confianza

$\underline{X}$ es una m.a. con $X_1 \sim F_\theta(x)$

$\hat{\theta}$ = est. puntual

**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 $\theta$ con nivel de confianza $1-\alpha$ será un conjunto tal que

$$\mathbb{P}(\theta \in S(\underline{X})) = 1 - \alpha. \ (*) \simeq 0{,}95$$

$\mathbb{R}$

$\theta^*$

Obs: $\theta$ **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. $\theta \in \Theta$

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

Estandarizar $\bar{X}$. Si $\underline{X}$ es una m.a. con $X_i \sim N(\mu, \sigma)$

$$\frac{\bar{X} - E[\bar{X}]}{\sqrt{Var(\bar{X})}} \xrightarrow[n\to\infty]{P} \underbrace{\frac{\bar{X} - \mu}{\sigma}\sqrt{n}}_{g(\underline{X}, \mu)} \sim \underbrace{N(0,1)}_{\text{No depende de } \mu}$$

**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 $\theta$. Sean $a$ y $b$ tales que $\mathbb{P}(a \leq U \leq b) = 1 - \alpha$. Luego,

$F_U^{-1}(\frac{\alpha}{2})$ $F_U^{-1}(1-\frac{\alpha}{2})$.

$$S(\underline{X}) = \{\theta : a < g(\underline{X}, \theta) \leq b\} = (g^{-1}(a), g^{-1}(b))$$

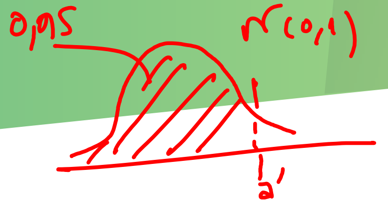
$P(X < q_3) = 0{,}75$

$F(q_3) = 0{,}75$

$q_3 = F^{-1}(0{,}75)$.

es una región de confianza para $\theta$. A $U$ se lo llama pivote.

# Ejercicio 4




con $X_1 \sim N(\mu, 4) \Rightarrow \sigma^2 = 4$

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

$P[\mu \in (a, +\infty)] = 0{,}95$

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

g es decreciente como función de $\mu$

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

$$P[\underbrace{g(\underline{X}, \mu)}_{= U \sim N(0,1)} < a'] = 0{,}95 \Rightarrow F_U(a') = 0{,}95 \Rightarrow a' = F_U^{-1}(0{,}95) = 1{,}64$$

$$P\left( g(\bar{x},\mu) < 1{,}64 \right) = 0{,}95$$

$$P\left( \frac{\bar{x}-\mu}{\sigma}\sqrt{n} < 1{,}64 \right) = 0{,}95$$

$$P\left( \mu > \underbrace{\bar{x} - 1{,}64\,\frac{\sigma}{\sqrt{n}}}_{a(\underline{x}) \,=\, 2{,}6} \right) = 0{,}95$$

$$\left[ P\left( \mu \in \underbrace{\left( \bar{x} - 1{,}64\,\frac{\sigma}{\sqrt{n}}\,,\; +\infty \right)}_{\text{es un i.c. de nivel } 0{,}95.} \right) = 0{,}95 \right]$$

# Algunos resultados importantes

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

$$Z = \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \sim \mathcal{N}(0,1) \quad \sigma \to S$$

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

TP1. $X \sim \mathcal{N}(0,1) \to X^2 \sim \chi^2_1$.

$\sum_{n=1}^{n} X_i^2 \sim \chi^2_n$

$V$ y $W$ son independientes

Si $\boxed{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, $U = \sqrt{n}\frac{(\bar{X}-\mu)}{S} \sim t_{n-1}$

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 $\boxed{\frac{X}{\sqrt{Y/n}} \sim t_n}$

# 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: $U(\underline{X}, \mu) = \frac{(\overline{X} - \mu)}{S}\sqrt{n} \sim t_{n-1} \xrightarrow[n \to \infty]{} \mathcal{N}(0,1)$
- Para el desvío con media conocida: $U(\underline{X}, \sigma) = \frac{\sum_{i-1}^{n}(X_i - \mu)^2}{\sigma} \sim \chi^2_n$
- 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

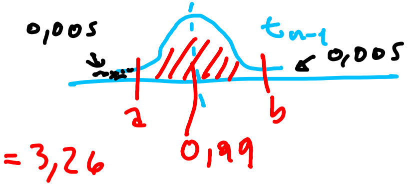
$X_i \sim N(\mu, \sigma^2)$

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, $\boxed{\mu = 2, \sigma = 3,}$ simular la muestra y calcular el IC resultante de la misma.

$U(\underline{X}, \mu) = \frac{\bar{X} - \mu}{S}\sqrt{n} \sim t_{n-1}$ con $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$P(a < U(\underline{X}, \mu) < b) = 0{,}99$




$a = F_U^{-1}(0{,}005) = -3{,}26$ y $b = F_U^{-1}(0{,}995) = 3{,}26$

$a < U < b$

$a < \frac{\bar{X} - \mu}{S} \sqrt{n} < b$

$\underbrace{\bar{X} - a \frac{S}{\sqrt{n}}}_{a(\underline{X})} > \mu > \underbrace{\bar{X} - b \frac{S}{\sqrt{n}}}_{b(\underline{X})}$ es un i.c. de nivel 0,99

$n = 50$, generamos muestra y vemos que $\bar{x} = 1{,}36$ y $S = 3{,}05$

$\Rightarrow (-0{,}16 \, , \, 2{,}65)$ es un i.c. de nivel 0,99 para $\mu$.

# Bibliografía


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