Probabilidad y estadística

Clase 5

- $F(x) = \mathbb{P}(X \leq x)$ "ADN" ✓
- $f(x)$ en variables continuas

Método paramétrico

Asumo $X \sim \mathcal{E}(\lambda)$

Estimo $\hat{\lambda}$

$$\hat{F}(x) = 1 - e^{-\hat{\lambda} x}$$

$$\hat{f}(x) = \hat{\lambda} e^{-\hat{\lambda} x}$$

Estimación no paramétrica de distribuciones

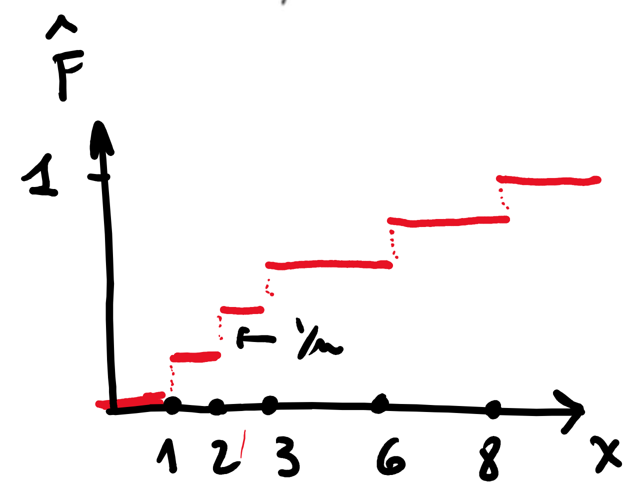
Función de distribución empírica

Partimos $(X_1, X_2, \dots, X_n) \sim F$

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.

50

$$\hat{P}(X \le x) = \widehat{F}_n(x) = \frac{\overbrace{\sum_{i=1}^{n} I\{X_i \le x\}}^{\#\{X_i \le x\}}}{n}$$



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 software.

Ejercicio 1

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

–7,9, –1,9, –0,9, ..., 25,5

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

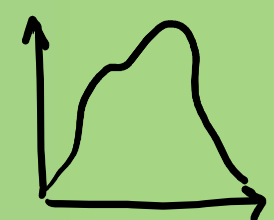
Estimación de densidad

- histograma
- kernels

X es v.a continua

$f_X(x) \geq 0$

$\int_{sop} f(x) = 1$



Histogramas

$$fr_{B_i} = \frac{\#\{x \in B_i\}}{n}$$

Se selecciona un origen x_0 y se divide la recta real en intervalos de longitud h

$$B_j = [x_0 + (j-1)h, x_0 + jh], j \in \mathbb{N}$$

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

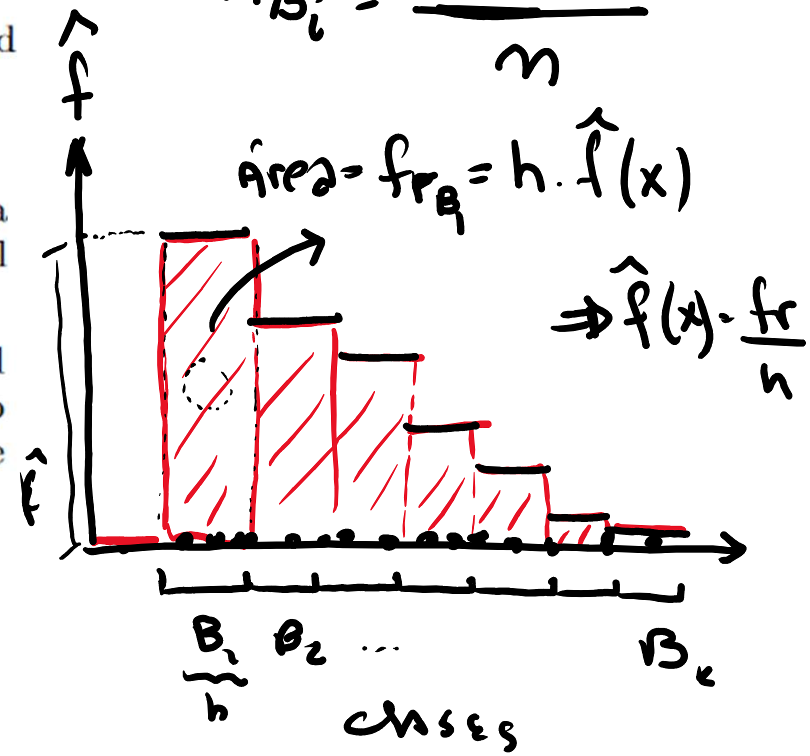
Para cada intervalo, se divide la frecuencia absoluta por la cantidad total de la muestra n (para convertirlas en frecuencias relativas, análogo a como se hace con las probabilidades) y por la longitud h (para asegurarse que el area debajo del histograma sea igual a 1):

Formalmente, el histograma está dado por:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j} \mathbf{1}(x_i \in B_j)\mathbf{1}(x \in B_j)$$

$$\frac{15}{160}$$

$$\sum_{i=1}^{n} \hat{f}(x)$$



Apunte de Histograma - PyE FIUBA

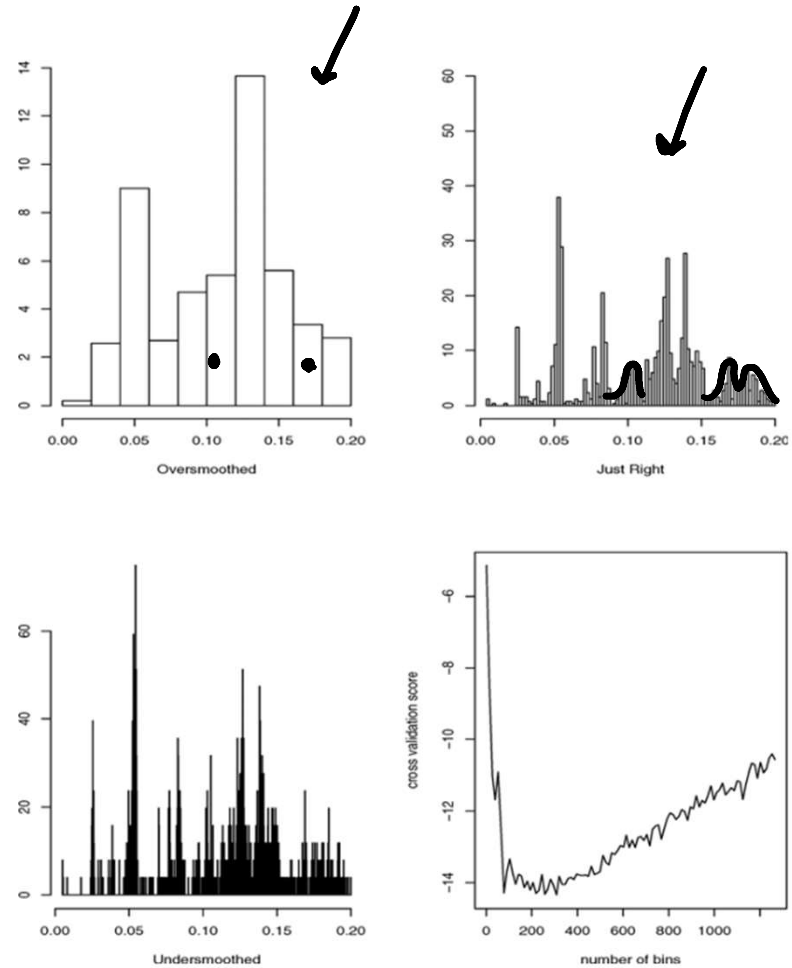
Ejercicio 2

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’ usando software.

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

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



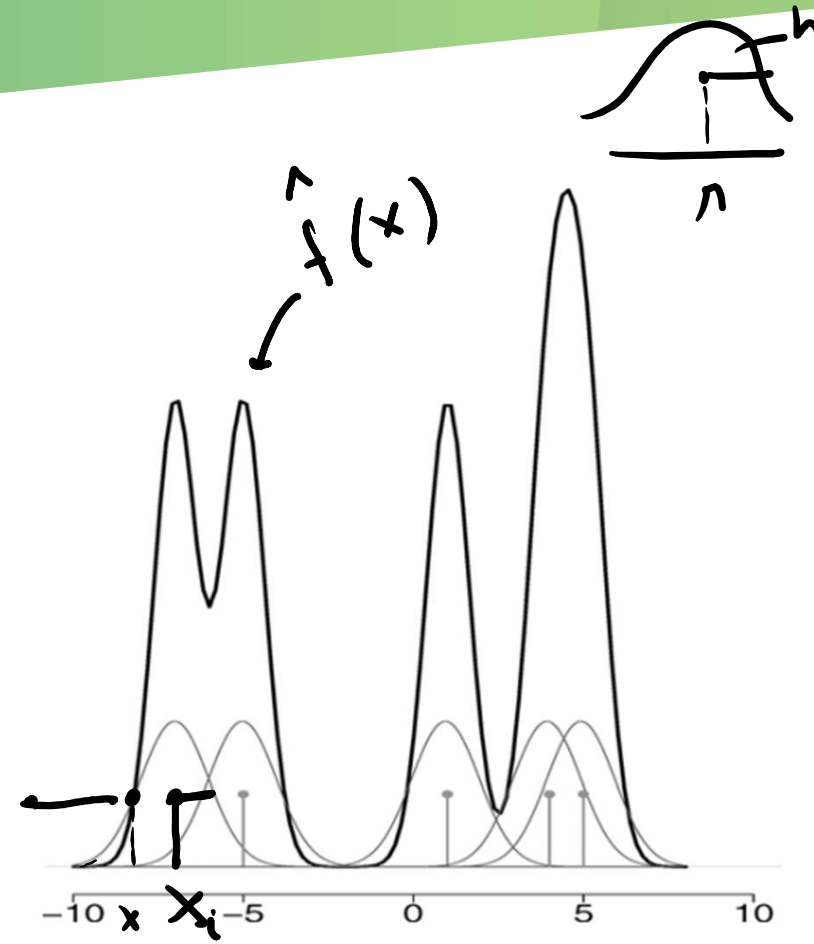
Estimación de densidad por kernel

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.

KDE - Ejemplo

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

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

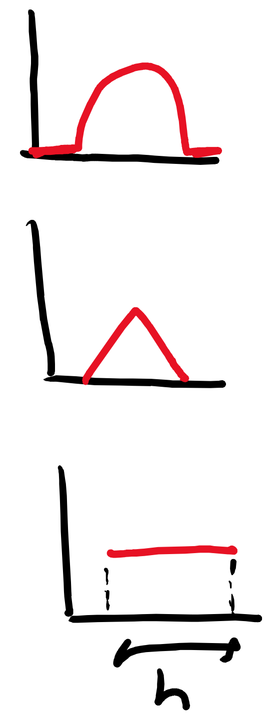


h : ventana de suavizado

Kernel – Definición

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

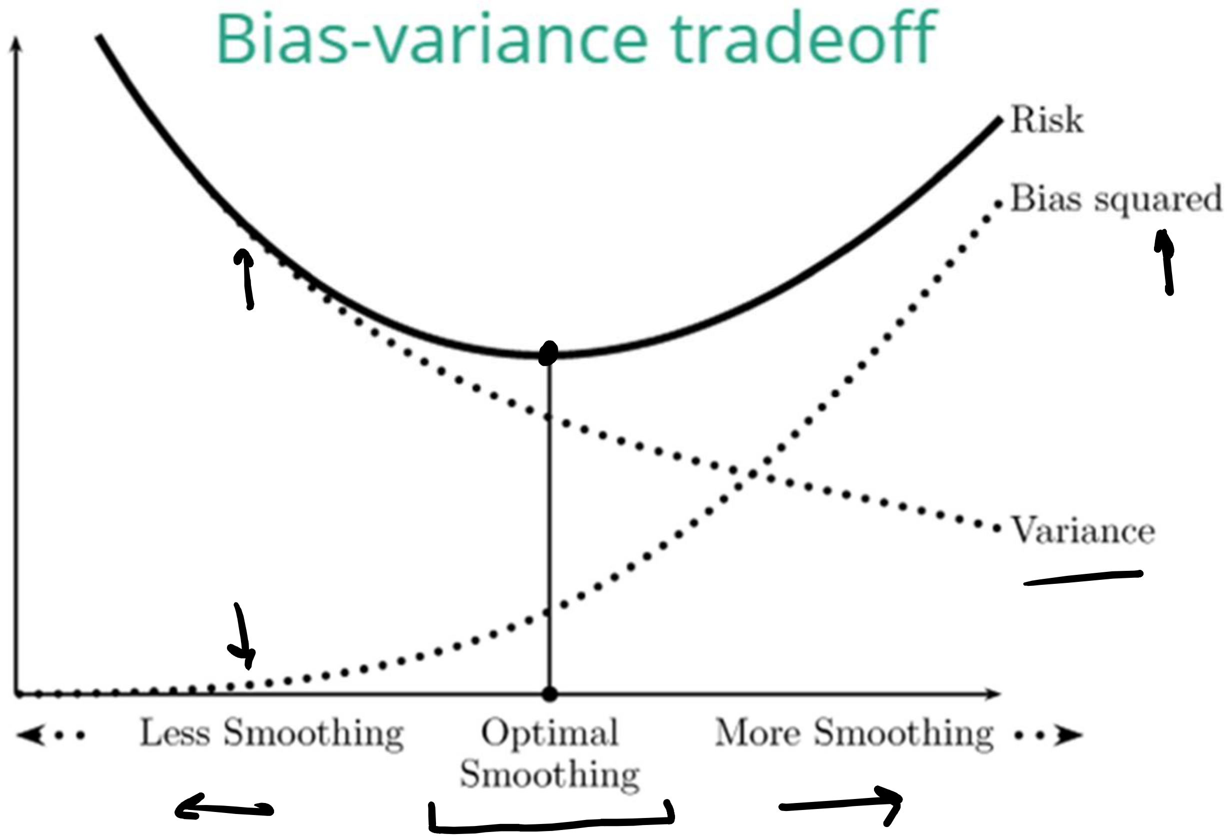
  Es óptima en el sentido de error cuadrático medio

- Gaussiano (simple)

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.

Bias-Variance Tradeoff



Intervalos de confianza

Estimación por intervalo.

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.

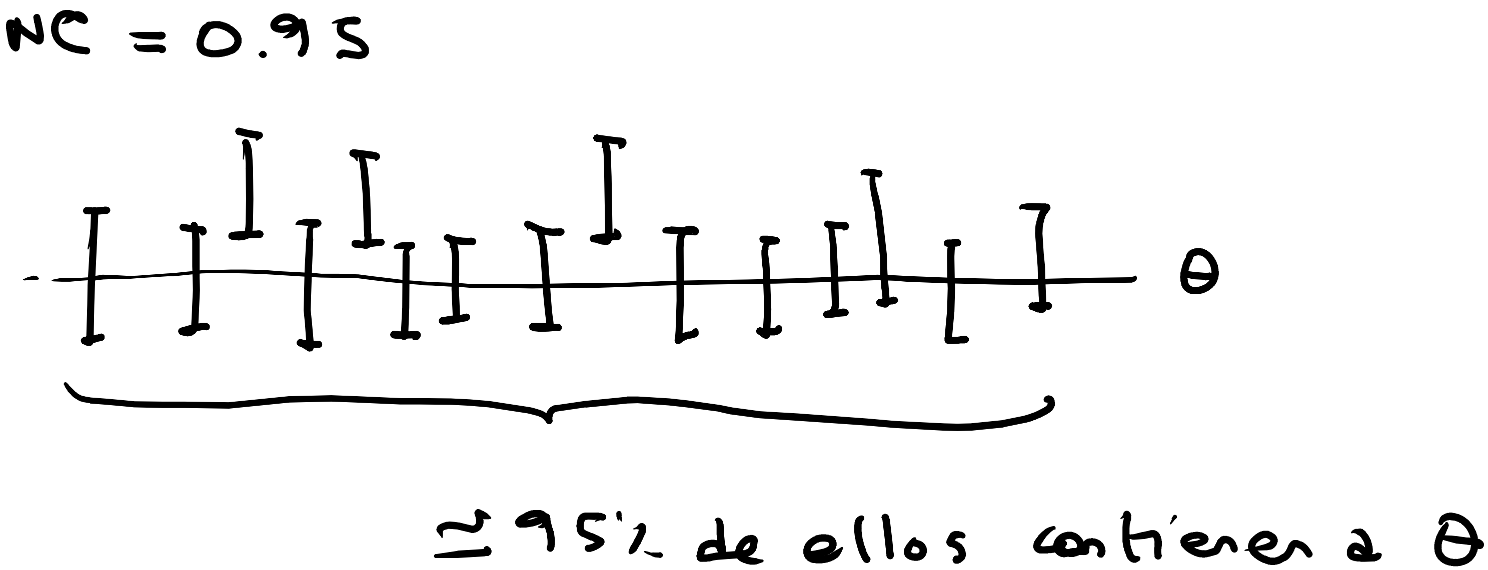
$$\underline{x} = (X_1 X_2 \dots X_n) \overset{iid}{\sim} F$$

$$\bullet)\quad \mathbb{P}\left(A(\underline{x}) \le \theta \le B(\underline{x})\right) = \underbrace{1 - \alpha}_{\text{nivel de confianza}}$$

error

$\simeq 1$

¿Qué es un IC?



En la siguiente api podemos visualizar un poco mejor qué es un IC con simulaciones.

Ejercicio 4

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.

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

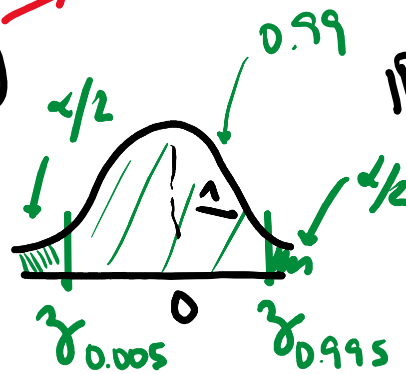
Ejercicio 4

$X \sim N(\mu, \sigma^2)$ CONOCIDA

OBJETIVO.

$\mathbb{P}(A(\underline{x}) \le \mu \le B(\underline{x})) = 0.99$

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$



$1-\alpha = 0.99 \qquad \alpha = 0.01$

$$\mathbb{P}\left(z_{\alpha/2} \le \boxed{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}} \le z_{1-\alpha/2}\right) = 1-\alpha$$

$$\mathbb{P}\Big(\underbrace{\bar{X} - z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}}_{A(\underline{x})} \le \mu \le \underbrace{\bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}}_{B(\underline{x})}\Big) = 1-\alpha$$

$$\bar{X} \pm z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$$

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 $\underbrace{S(\underline{X})}_{\text{REGIÓN}}$ para θ con nivel de confianza $1-\alpha$ será un conjunto tal que

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

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.

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 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 varianza de la población

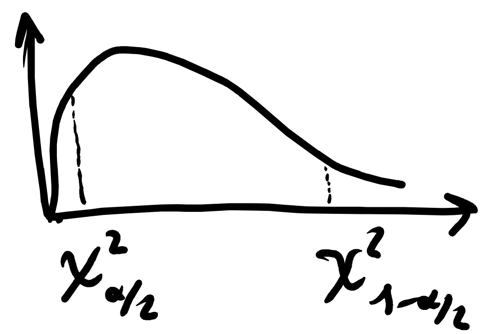
Ejercicio 5

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

$$\mathbb{P}\left(a(\underline{X}) \le \sigma^2 \le b(\underline{X})\right) = 1-\alpha$$

ESTIMADOR PARA $\sigma^2 \longrightarrow S^2 = \dfrac{\sum (X_i - \bar{X})^2}{n-1}$

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{\underbrace{n-1}_{\text{PARÁMETRO}}}$$



$$\mathbb{P}\left(\chi^2_{n-1,\alpha/2} \le (n-1)\frac{S^2}{\sigma^2} \le \chi^2_{1-\alpha/2}\right) = 1-\alpha$$

$$\mathbb{P}\left(\overbrace{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}^{a(\underline{X})} \le \sigma^2 \le \overbrace{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}}^{b(\underline{X})}\right) = 1-\alpha$$

Ejercicio 5

simulación $n = 50$ $s^2 = 2,4213$

quiero IC del 99% $1-\alpha = 0.99$ $\alpha = 0.01$

$1-\alpha/2 = 0.995$ $\alpha/2 = 0.005$

$n-1 = 49$

$\chi^2_{49,0.995} = 78,23$

$\chi^2_{49,0.005} = 27,24$

$$a(\underline{x}_{obs}) = \frac{49 \cdot 2,4213}{78,23} = 1,516$$

$$b(\underline{x}_{obs}) = \frac{49 \cdot 2,4213}{27,24} = 4,35$$

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