

Función de distribución empírica

Teorema fundamental de la estadística (Glivenco y Canteli)

Inferencia Estadística

6 de octubre de 2021

Notación

X población; variable aleatoria $X : \Omega \longrightarrow \mathbb{R}$.

X_1, \dots, X_n, \dots sucesión de variables aleatorias independientes e idénticamente distribuidas como X

$F(\cdot)$ función de distribución de X

$F_n(\cdot)$ función de distribución empírica de X_1, \dots, X_n

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(X_i)$$

Notación

$\xrightarrow[n \rightarrow \infty]{\text{c.s.}}$ convergencia casi seguro cuando n tiende a infinito:

$$\begin{aligned} A_1, \dots, A_n, \dots: \Omega &\longrightarrow \mathbb{R}^\infty \\ \omega &\longmapsto (a_1, \dots, a_n, \dots) \end{aligned}$$

$$\begin{aligned} A_n \xrightarrow[n \rightarrow \infty]{\text{c.s.}} a &\iff \Pr\left[\lim_n A_n = a\right] = 1 \\ &\iff \Pr\left[\left\{\omega \mid \lim_n a_n = a\right\}\right] = 1 \end{aligned}$$

Función de distribución empírica

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(X_i)$$

$$n \cdot F_n(x) \sim \mathcal{B}(n, \Pr[X \leq x]) = \mathcal{B}(n, F(x))$$

Ley fuerte de los grandes números: $\forall x \in \mathbb{R}, F_n(x) \xrightarrow[n \rightarrow \infty]{\text{c.s.}} F(x)$

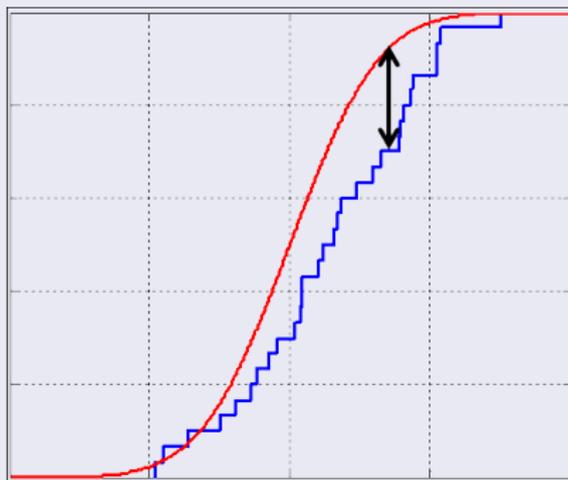
$$\text{T.C.L.: } \forall x \in \mathbb{R}, \frac{F_n(x) - F(x)}{\sqrt{\frac{F(x)[1-F(x)]}{n}}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

Lo mismo ocurre con $F_n(x^-) = \lim_{t \rightarrow x^-} F_n(t)$ y $F(x^-) = \Pr[X < x]$.

Notación

Δ_n distancia máxima en vertical entre dos funciones:

$$\Delta_n = D_\infty(F_n, F) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$



Funciones de distribución

```
D <- "unif" # norm exp ...
X <- get (paste0 ("r", D)) (10) # 100 1000 ...
F <- get (paste0 ("p", D)) # teórica
Fn <- ecdf (X) # empírica
plot (Fn)
plot (F, min(X), max(X), col=2, add=TRUE)
```

Distancia máxima

```
X <- sort (X) ; n <- length (X)
Dn1 <- abs (F(X) - Fn(X))
Dn2 <- abs (F(X) - c(0,Fn(X[-n])))
i1 <- which.max (Dn1) ; i2 <- which.max (Dn2)
I <- if (Dn1[i1] > Dn2[i2]) 1 else 2
i <- c(i1,i2)[I] ; Xi <- X[i]
Yi <- c (Fn(Xi), c(0,Fn(X))[i]) [I]
arrows (Xi, F(Xi), Xi, Yi, code=3, col=3)
```

Teorema (fundamental de la estadística o de Glivenco y Canteli)

$$\Delta_n \xrightarrow[n \rightarrow \infty]{\text{c.s.}} 0$$

Lema

Sean A_1, A_2, \dots sucesos con $\Pr[A_i] = 1 \forall i$
Entonces $\Pr[\bigcap_i A_i] = 1$.

Demostración.

$$\Pr\left[\bigcap_i A_i\right] = 1 - \Pr\left[\bigcup_i A_i^c\right] \geq 1 - \sum_i \Pr[A_i^c] = 1 - \sum_i 0 = 1$$

Teorema (fundamental de la estadística o de Glivencov y Canteli)

$$\Delta_n \xrightarrow[n \rightarrow \infty]{\text{c.s.}} 0$$

Demostración.

Veremos dos situaciones:

- 1 variable discreta con número finito de valores
- 2 caso general

En cada situación se consideran

- sucesión de variables (X_1, \dots, X_n, \dots)
- sucesión particular $\omega = (x_1, \dots, x_n, \dots)$

Caso discreto finito

X toma valores $x_j, j \in \{1, \dots, k\}$, con $x_j < x_{j+1}$.

$$F \text{ y } F_n \text{ escalonadas} \implies \Delta_n(\omega) = \max_j |F_n(x_j, \omega) - F(x_j)|$$

$\forall j$, sea $A_j = \{\omega \mid \lim F_n(x_j, \omega) = F(x_j)\}$. Sea $A = \bigcap_{j=1}^k A_j$.

$$F_n(x_j) \xrightarrow[n \rightarrow \infty]{\text{c.s.}} F(x_j) \implies \Pr[A_j] = 1 \implies \Pr[A] = 1$$

$$\omega \in A \implies \omega \in A_j \implies \forall \epsilon \exists n_{j,\epsilon} \forall n > n_{j,\epsilon}, |F_n(x_j, \omega) - F(x_j)| < \epsilon$$

$$\omega \in A \implies \forall \epsilon \exists n_\epsilon = \max_{j=1}^n n_{j,\epsilon} \forall n > n_\epsilon, \Delta_n(\omega) < \epsilon$$

$$\implies \Pr[\lim \Delta_n = 0] = 1 \implies \Delta_n \xrightarrow[n \rightarrow \infty]{\text{c.s.}} 0$$

Caso general

Sea $k \in \mathbb{N}$ y considérense los cuantiles

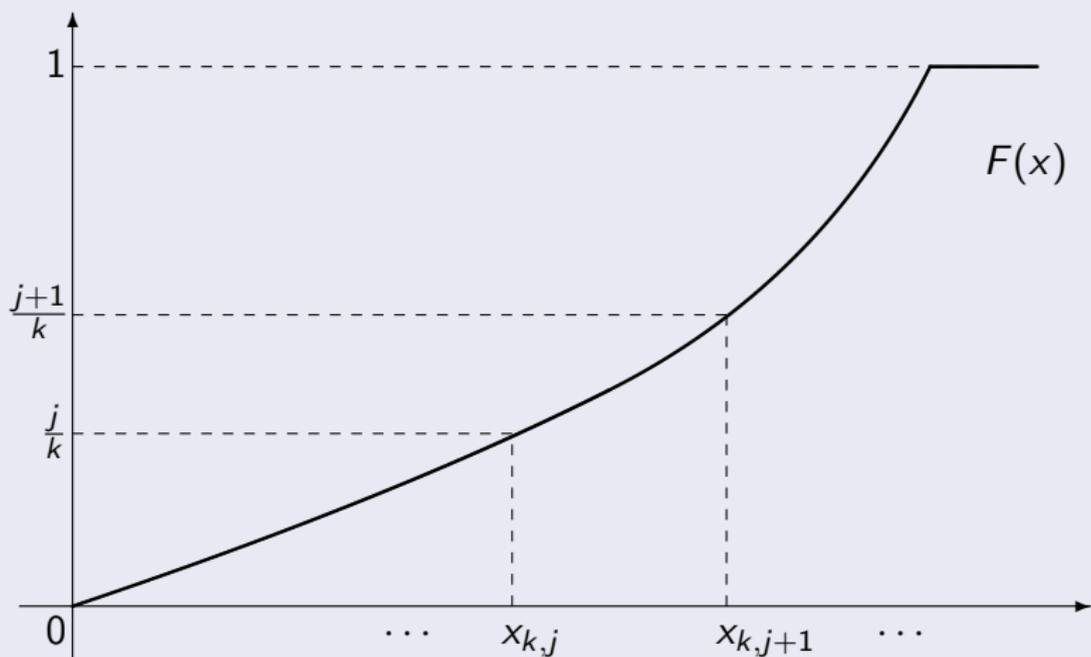
$$x_{k,j} = \inf \left\{ x \mid F(x) \geq \frac{j}{k} \right\} \quad j = 1, \dots, k$$

Más adelante se usará que

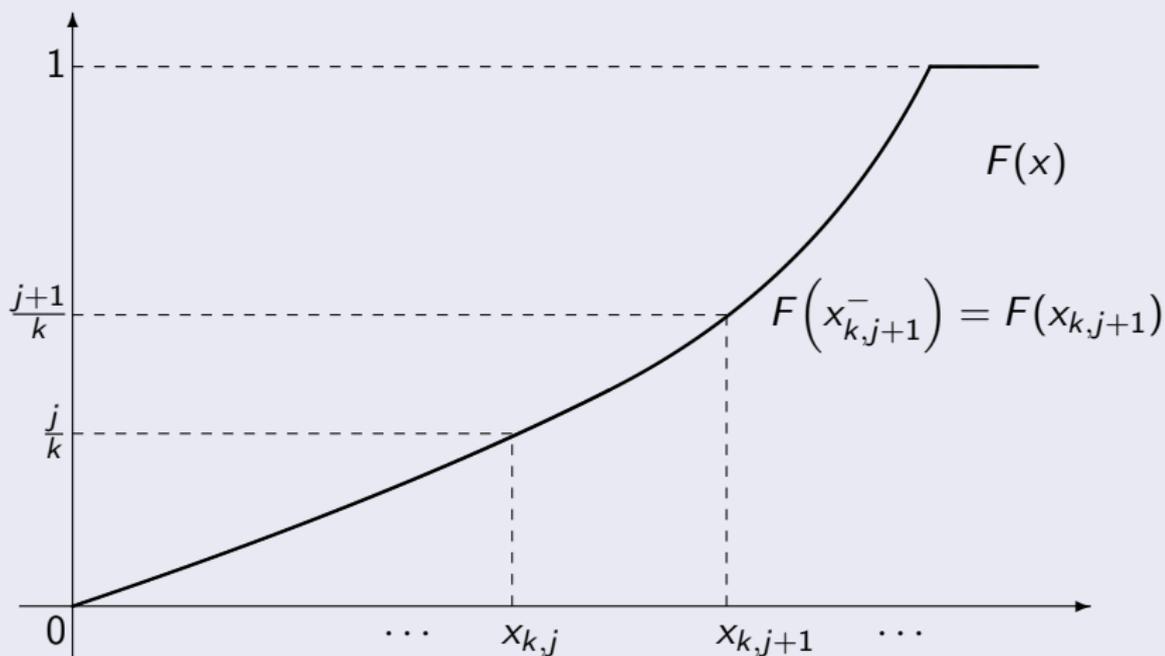
$$F(x_{k,j+1}^-) \leq F(x_{k,j}) + \frac{1}{k}$$

Se verá ahora la interpretación gráfica de esa fórmula para ilustrar los $x_{k,j}$.

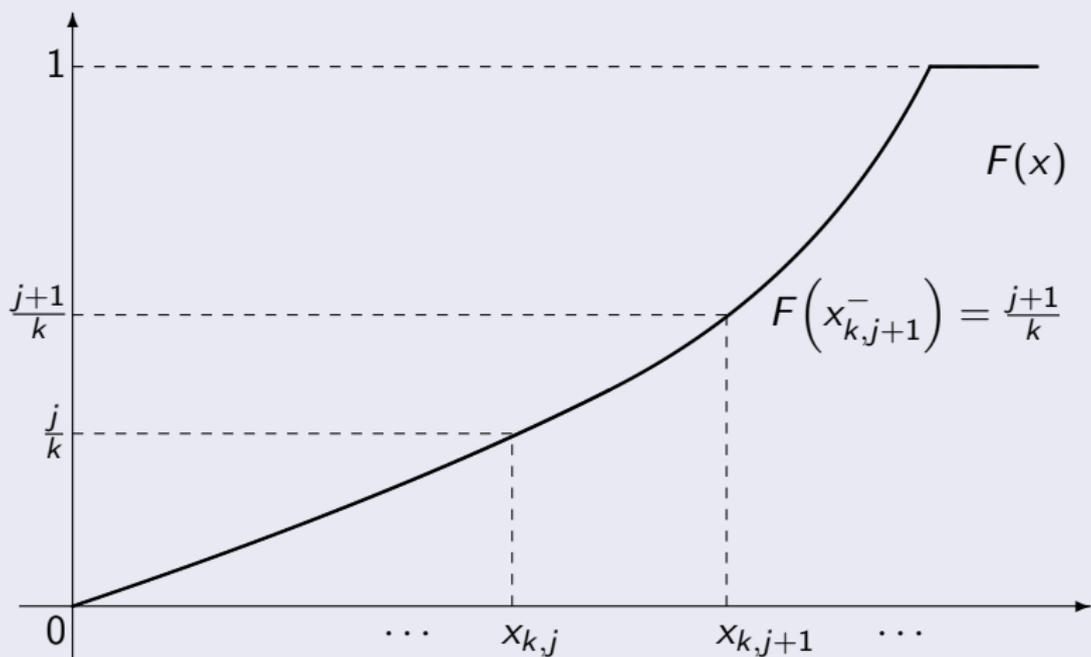
Caso continuo.



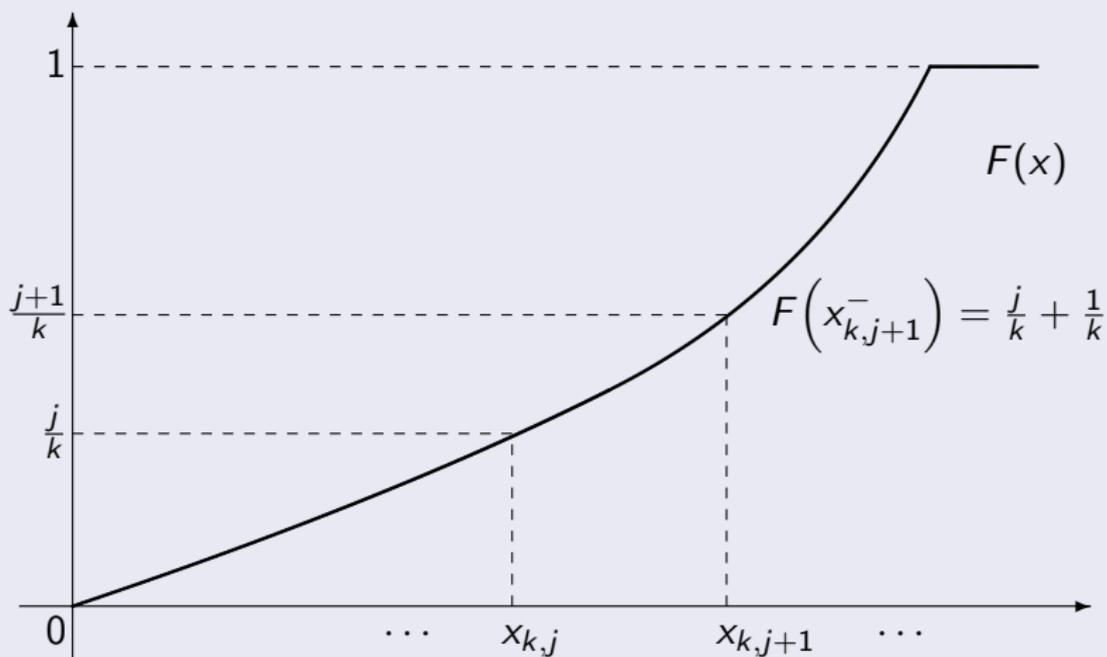
Caso continuo.



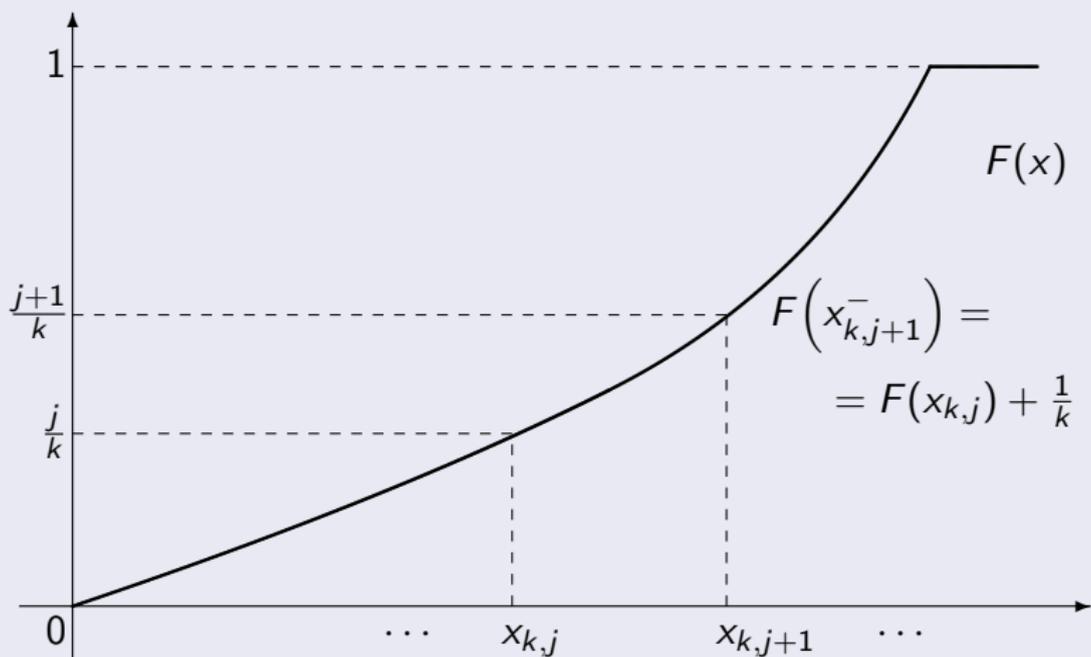
Caso continuo.



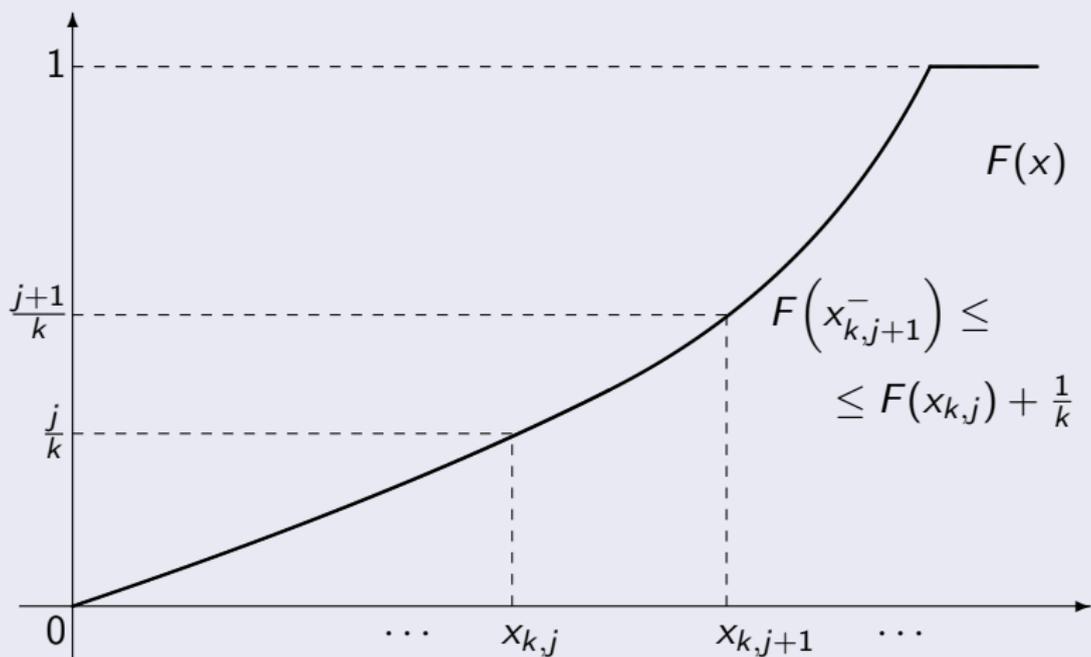
Caso continuo.



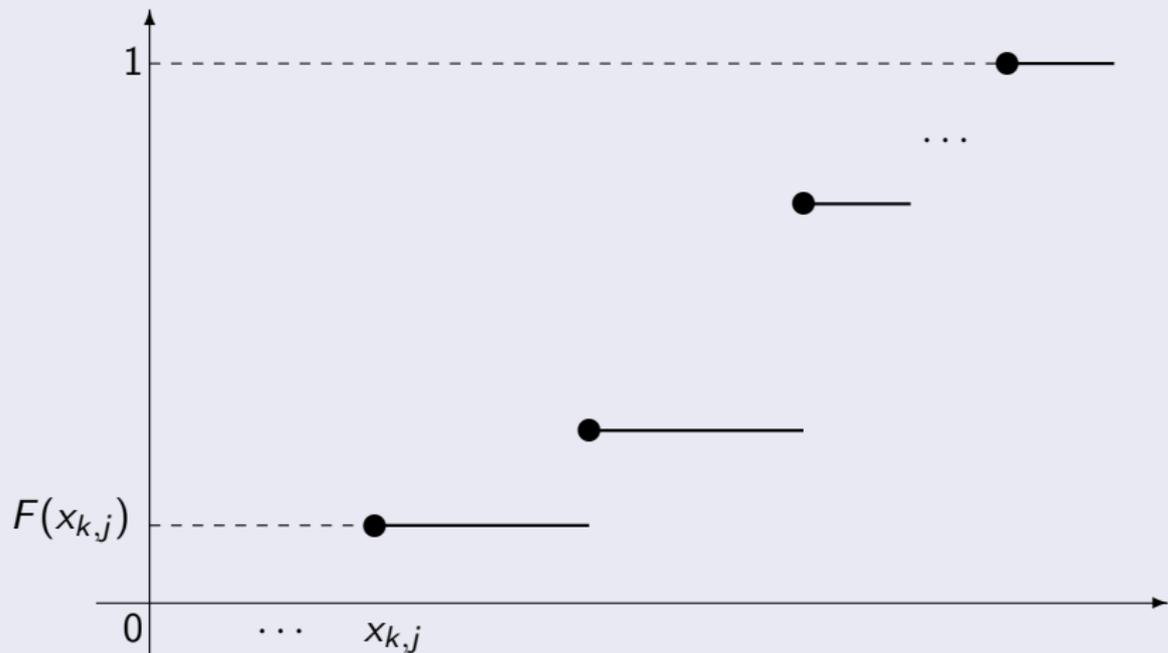
Caso continuo.



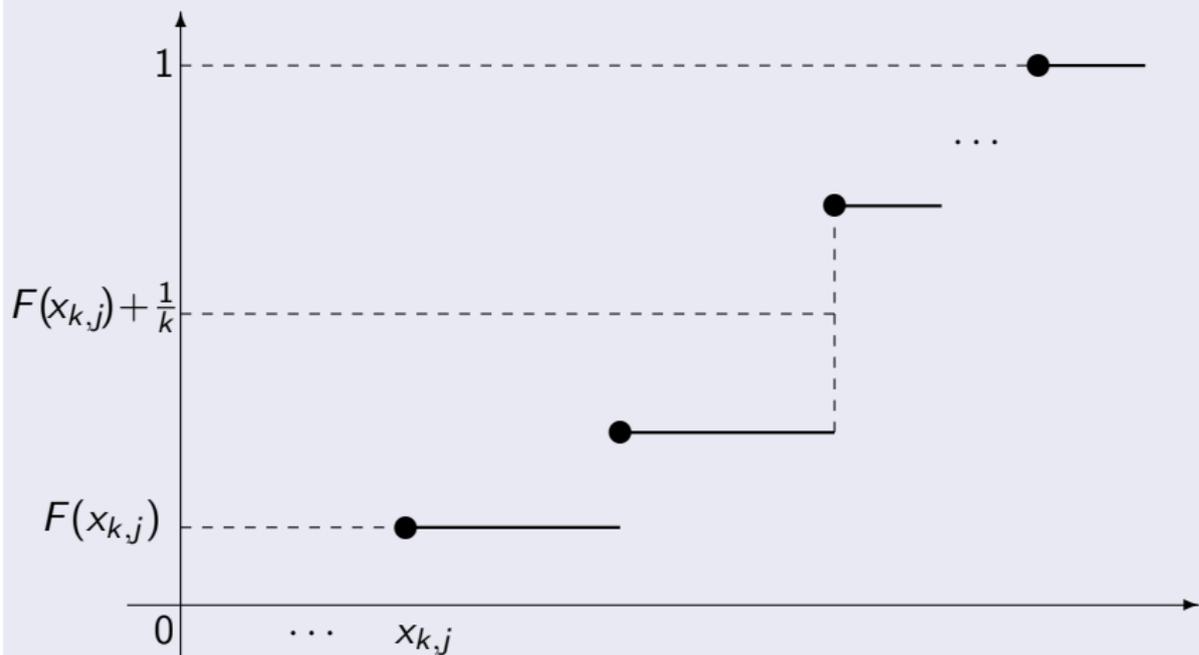
Caso continuo.



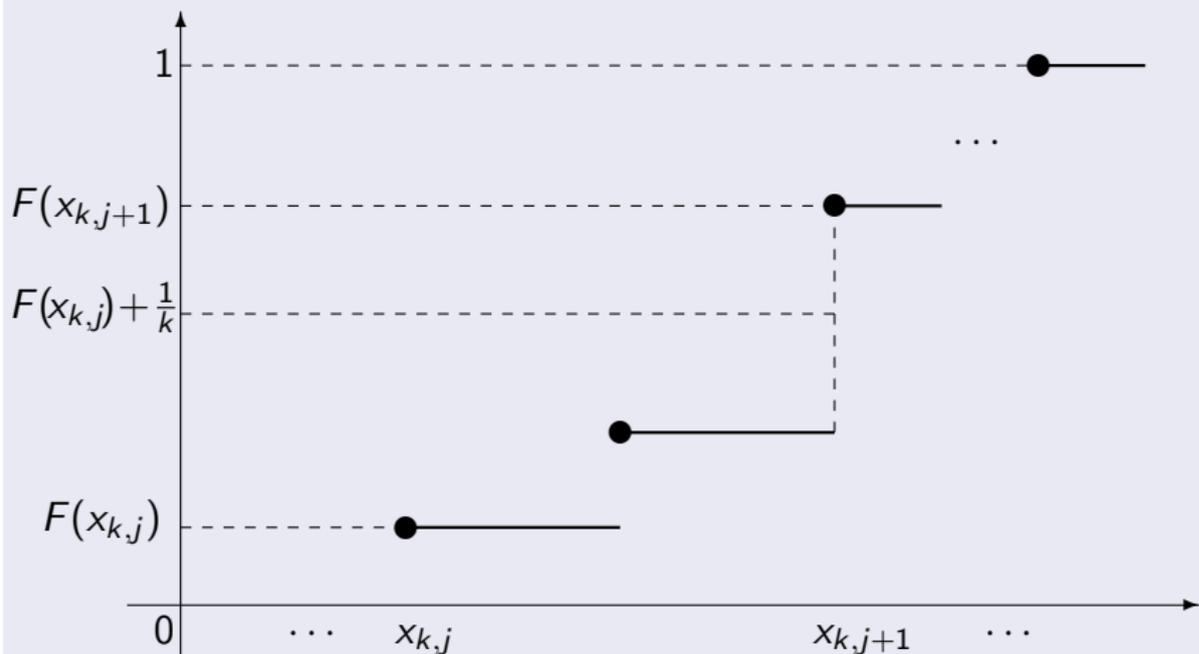
Caso discreto (sin coincidencia).



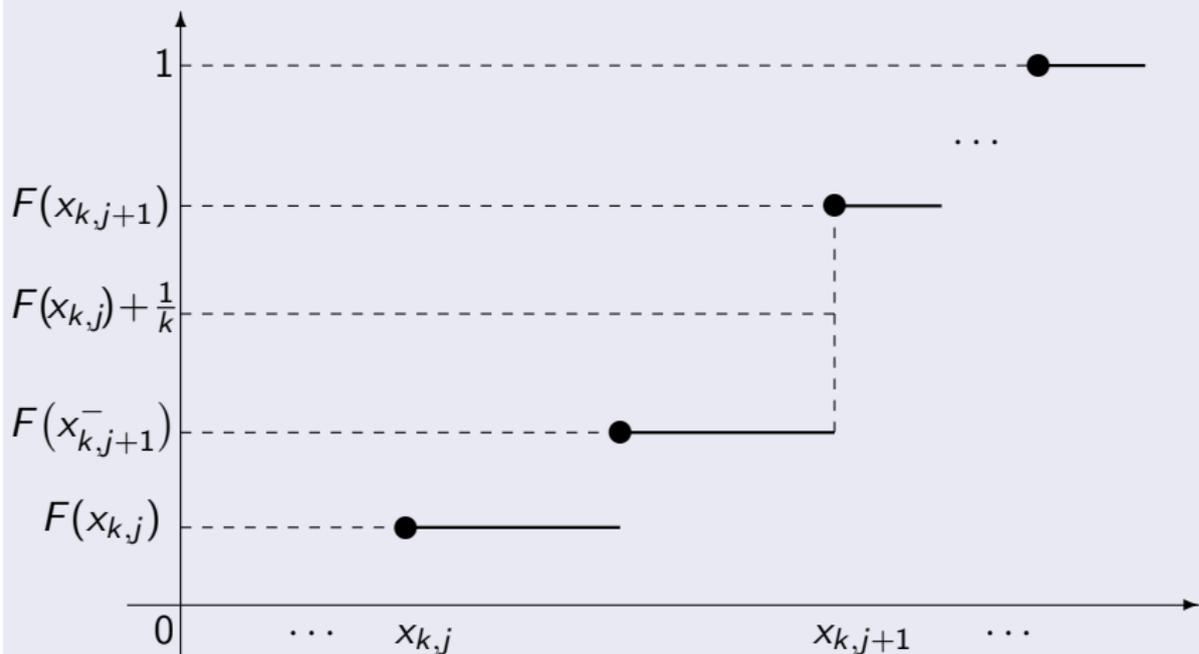
Caso discreto (sin coincidencia).



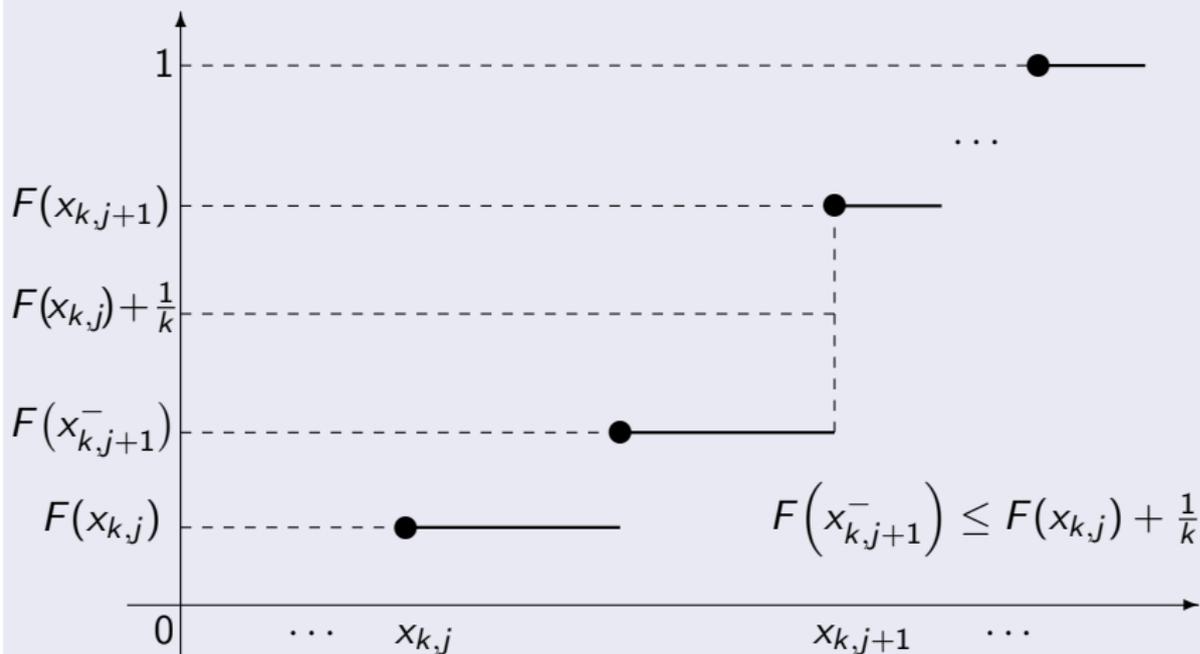
Caso discreto (sin coincidencia).



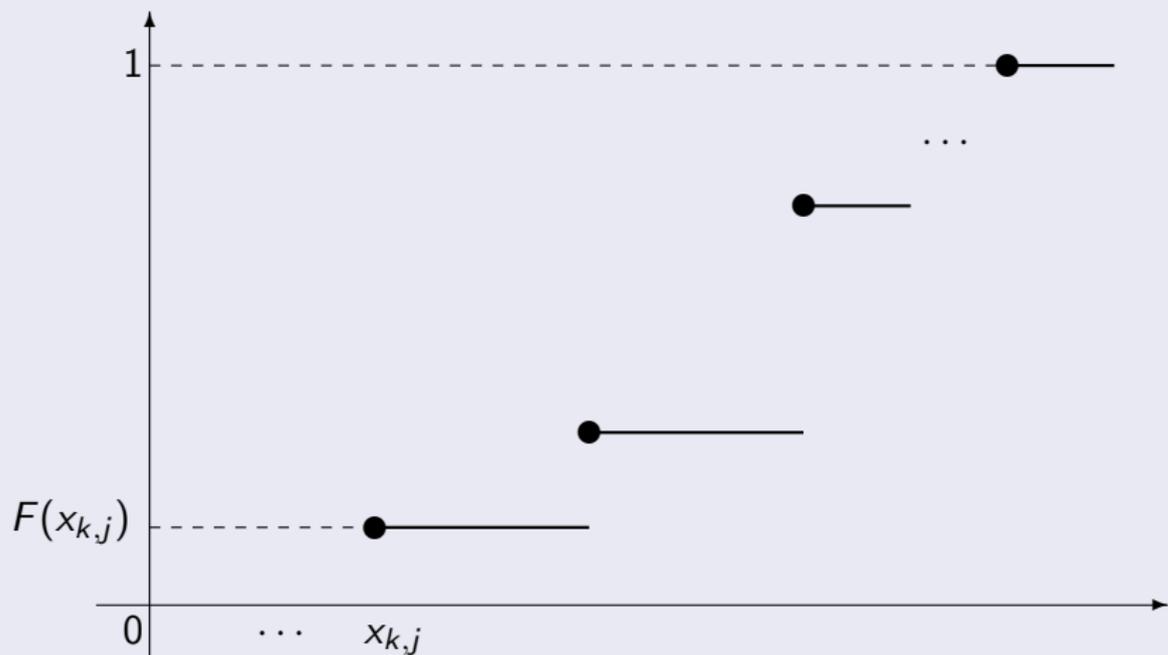
Caso discreto (sin coincidencia).



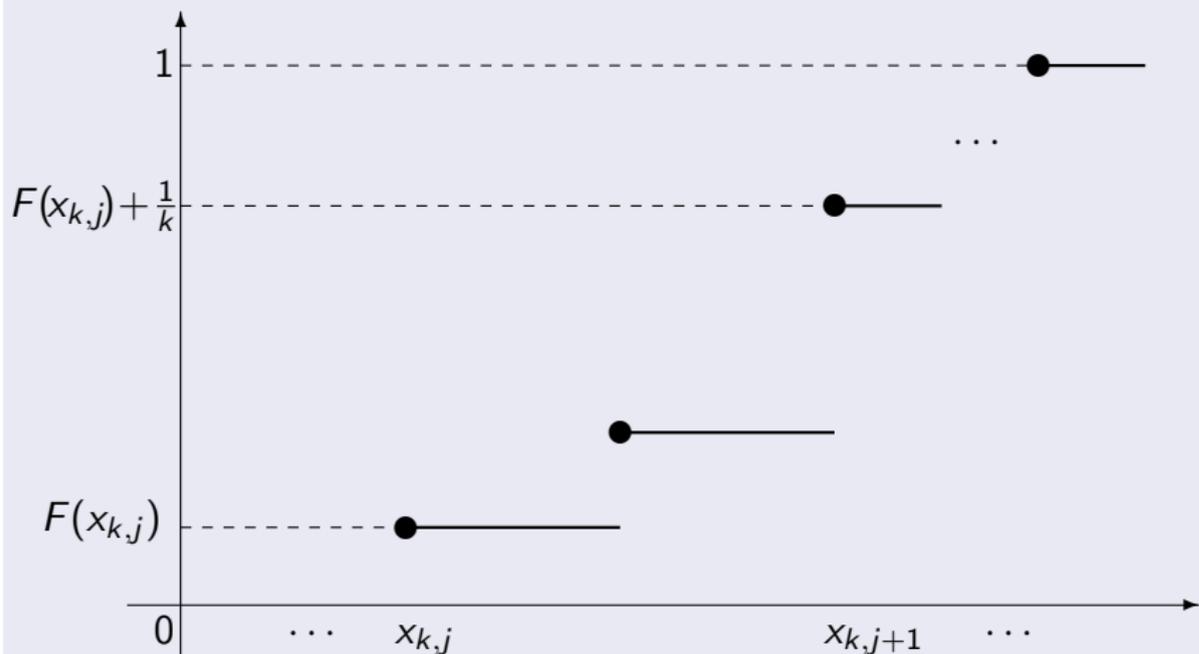
Caso discreto (sin coincidencia).



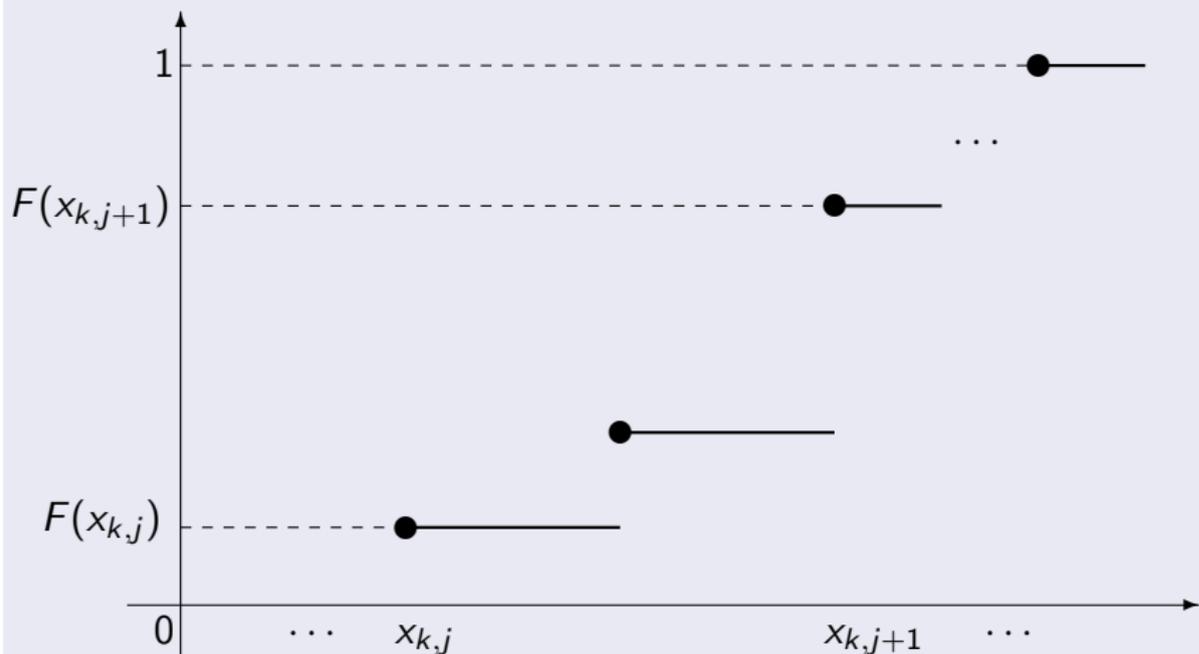
Caso discreto (con coincidencia).



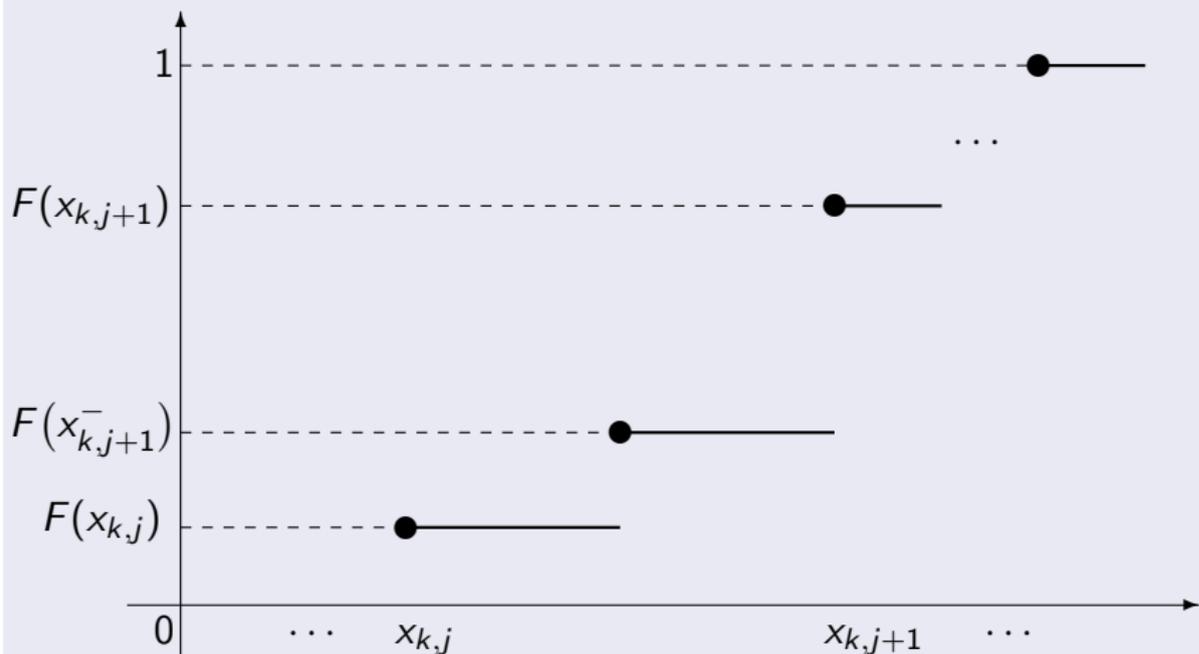
Caso discreto (con coincidencia).



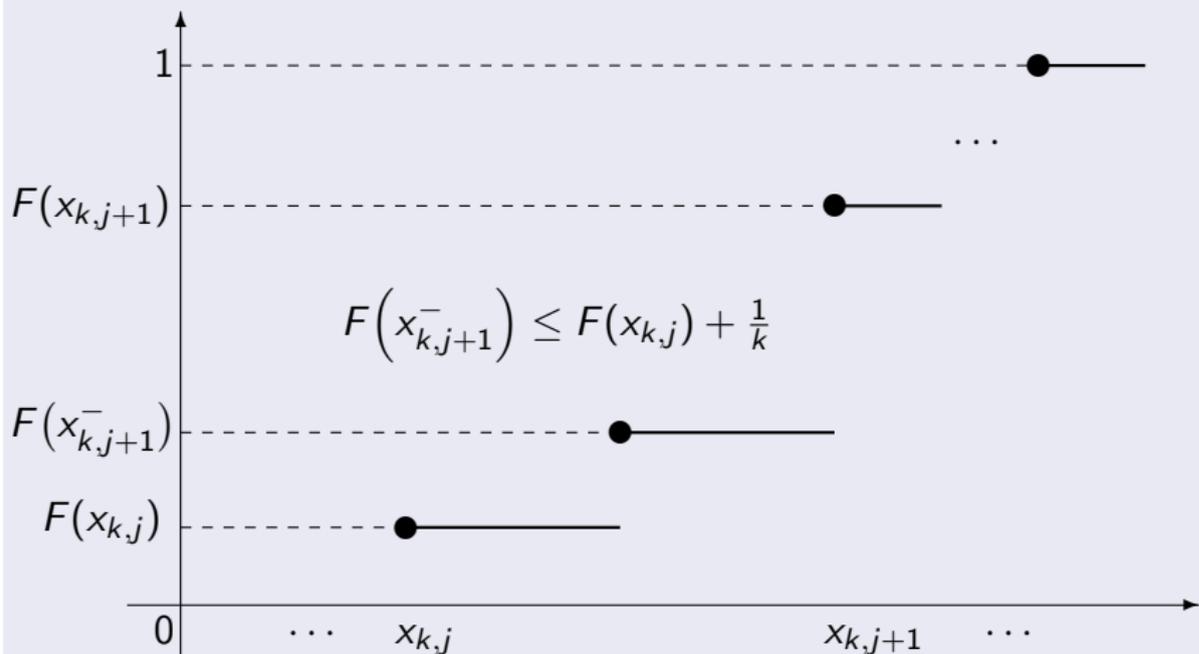
Caso discreto (con coincidencia).



Caso discreto (con coincidencia).



Caso discreto (con coincidencia).



Sabemos que $\forall k \in \mathbb{N} \quad \forall 1 \leq j \leq k$

$$F_n(x_{k,j}) \xrightarrow[n \rightarrow \infty]{\text{c.s.}} F(x_{k,j}) \quad \text{y} \quad F_n(x_{k,j}^-) \xrightarrow[n \rightarrow \infty]{\text{c.s.}} F(x_{k,j}^-)$$

Sean

$$A_{k,j} = \{\omega \mid \lim F_n(x_{k,j}, \omega) = F(x_{k,j})\}$$

$$B_{k,j} = \{\omega \mid \lim F_n(x_{k,j}^-, \omega) = F(x_{k,j}^-)\}$$

$$C_k = \bigcap_{j=1}^k A_{k,j} \cap B_{k,j} \quad C = \bigcap_{k \in \mathbb{N}} C_k$$

Entonces

$$\forall k \forall j, \Pr[A_{k,j}] = \Pr[B_{k,j}] = \Pr[C_k] = \Pr[C] = 1$$

Para cada $\omega = (x_n)_{n \in \mathbb{N}} \in \Omega$ sea F_n la función de distribución empírica asociada a (x_1, \dots, x_n) y sea

$$\delta_n^k = \max_j \left\{ |F_n(x_{k,j}) - F(x_{k,j})|, |F_n(x_{k,j}^-) - F(x_{k,j}^-)| \right\}$$

luego

$$\forall \omega \in C, \lim_{n \rightarrow \infty} \delta_n^k = 0$$

Si $x \in [x_{k,j}, x_{k,j+1})$ entonces

$$\begin{aligned} F_n(x_{k,j}) &\leq F_n(x) \leq F_n(x_{k,j+1}^-) \\ \frac{j}{k} &\leq F(x_{k,j}) \leq F(x) \leq F(x_{k,j+1}^-) \leq \frac{j+1}{k} \end{aligned}$$

luego

$$F_n(x_{k,j}) - F(x_{k,j+1}^-) \leq F_n(x) - F(x) \leq F_n(x_{k,j+1}^-) - F(x_{k,j})$$

$$F_n(x_{k,j}) - F(x_{k,j+1}^-) \leq F_n(x) - F(x) \leq F_n(x_{k,j+1}^-) - F(x_{k,j})$$

$$\frac{j}{k} \leq F(x_{k,j}) \leq F(x_{k,j+1}^-) \leq \frac{j+1}{k} \implies F(x_{k,j+1}^-) \leq F(x_{k,j}) + \frac{1}{k}$$

entonces

$$F_n(x_{k,j+1}^-) - F(x_{k,j}) \leq F_n(x_{k,j+1}^-) + \frac{1}{k} - F(x_{k,j+1}^-) \leq \delta_n^k + \frac{1}{k}$$

$$F_n(x_{k,j}) - F(x_{k,j+1}^-) \geq F_n(x_{k,j}) - F(x_{k,j}) - \frac{1}{k} \geq -\delta_n^k - \frac{1}{k}$$

$$-\delta_n^k - \frac{1}{k} \leq F_n(x) - F(x) \leq \delta_n^k + \frac{1}{k}$$

Por tanto, $\forall k \in \mathbb{N} \forall j \in \{1, \dots, k-1\} \forall x \in [x_{k,j}, x_{k,j+1})$

$$|F_n(x) - F(x)| \leq \delta_n^k + \frac{1}{k}$$

Queda ver qué ocurre en los extremos,

- $x < x_{k,1}$
- $x \geq x_{k,k}$

Si $x < x_{k,1}$ entonces $F(x_{k,1}^-) \leq \frac{1}{k}$ y

$$\left. \begin{array}{l} 0 \leq F_n(x) \leq F_n(x_{k,1}^-) \\ 0 \leq F(x) \leq F(x_{k,1}^-) \end{array} \right\} \implies -F(x_{k,1}^-) \leq F_n(x) - F(x) \leq F_n(x_{k,1}^-)$$

Por definición de δ_n^k ,

$$F_n(x_{k,1}^-) \leq F(x_{k,1}^-) + \delta_n^k \leq \frac{1}{k} + \delta_n^k$$

$$-F(x_{k,1}^-) \geq -\frac{1}{k} \geq -\frac{1}{k} - \delta_n^k$$

$$\implies |F_n(x) - F(x)| \leq \frac{1}{k} + \delta_n^k$$

Si $x \geq x_{k,k}$ entonces $F(x_{k,k}) = 1$ y $F_n(x_{k,k}) = 1$,
luego $|F_n(x) - F(x)| = |1 - 1| = 0$.

Por tanto, $\forall x \in \mathbb{R}$

$$|F_n(x) - F(x)| \leq \frac{1}{k} + \delta_n^k$$

luego $\forall \omega \in \mathcal{C}$

$$\Delta_n(\omega) = \sup_{x \in \mathbb{R}} |F_n(x, \omega) - F(x)| \leq \frac{1}{k} + \delta_n^k$$

y

$$\lim_{n \rightarrow \infty} \delta_n^k = 0$$

Para todo $\omega \in C$

$$0 \leq \lim_{n \rightarrow \infty} \Delta_n(\omega) \leq \lim_n \frac{1}{k} + \delta_n^k = \frac{1}{k} + \lim_n \delta_n^k = \frac{1}{k}$$

Como eso se verifica $\forall k \in \mathbb{N}$,

$$\lim_{n \rightarrow \infty} \Delta_n(\omega) = 0 \quad \forall \omega \in C$$

y como $\Pr[C] = 1$ entonces

$$\Delta_n \xrightarrow[n \rightarrow \infty]{\text{c.s.}} 0$$

Interpretación

- $F(x)$ está dentro de la banda $F_n(x) \pm \epsilon$ con probabilidad 1.
- Se puede estimar F con precisión arbitraria siempre que se disponga de suficiente tamaño muestral.