# Tema 2. Estimación de la función de distribución

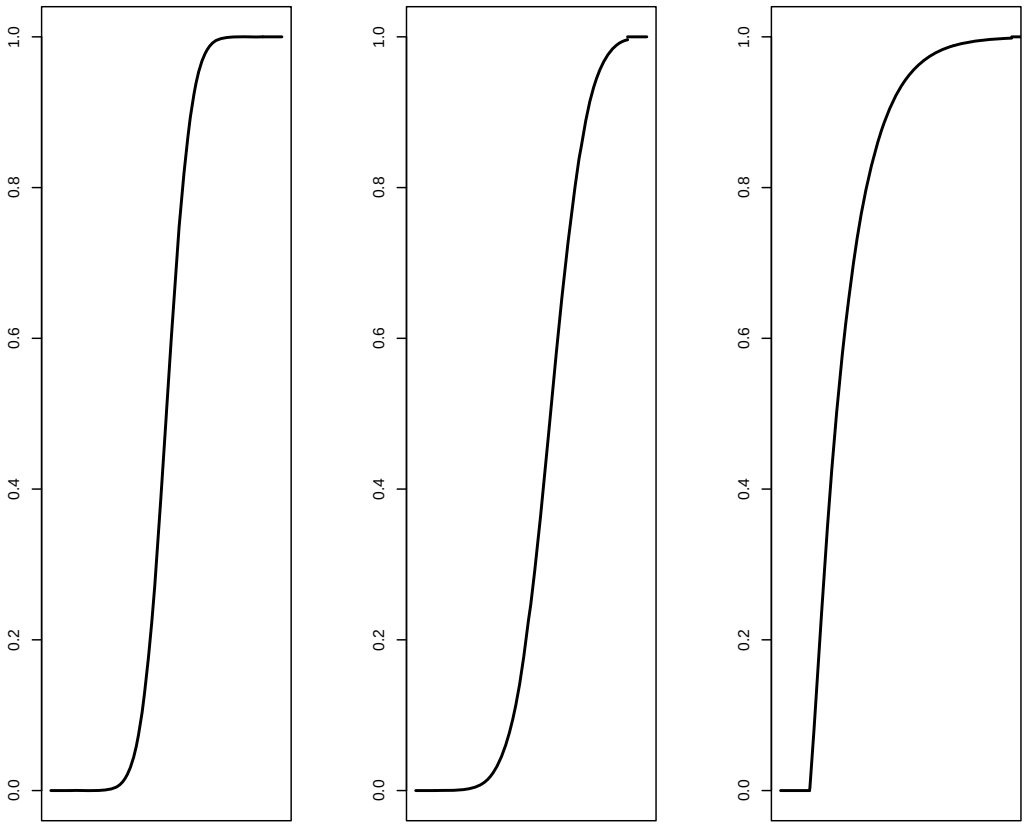
Rosa M. Crujeiras
Alberto Rodríguez

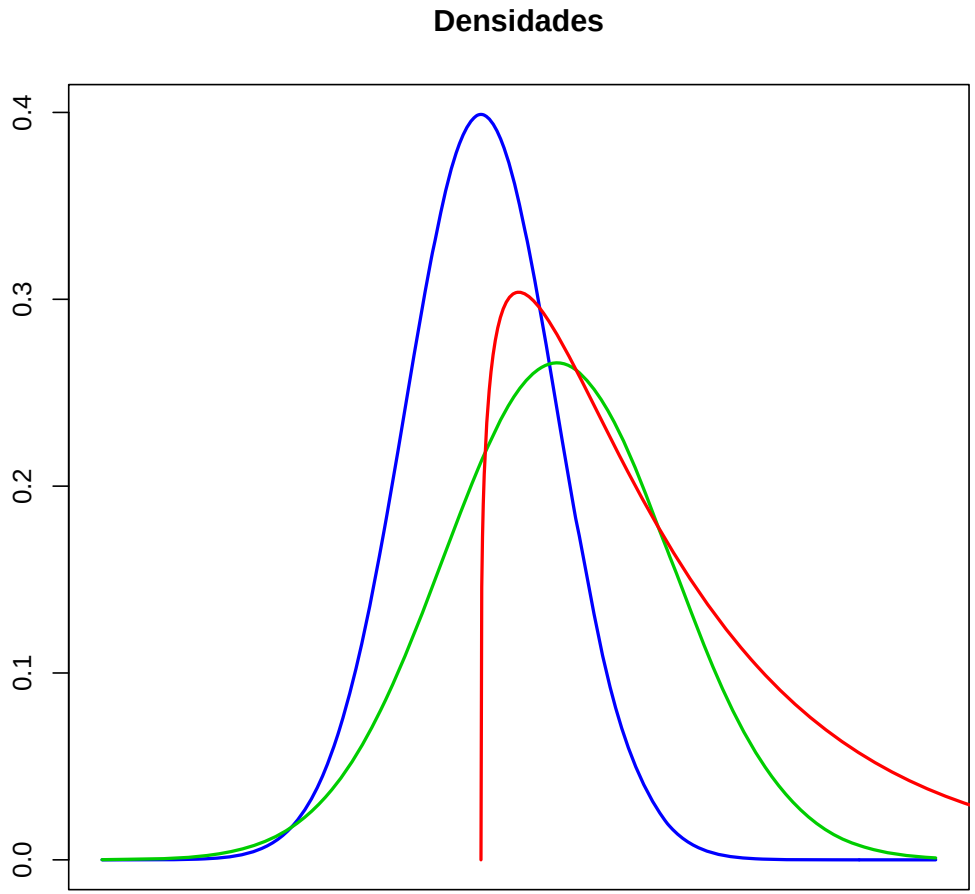

Dpto. de Estadística e Investigación Operativa
Máster en Técnicas Estadísticas
Curso 2009-2010

$X$ v.a. con distribución $F$:

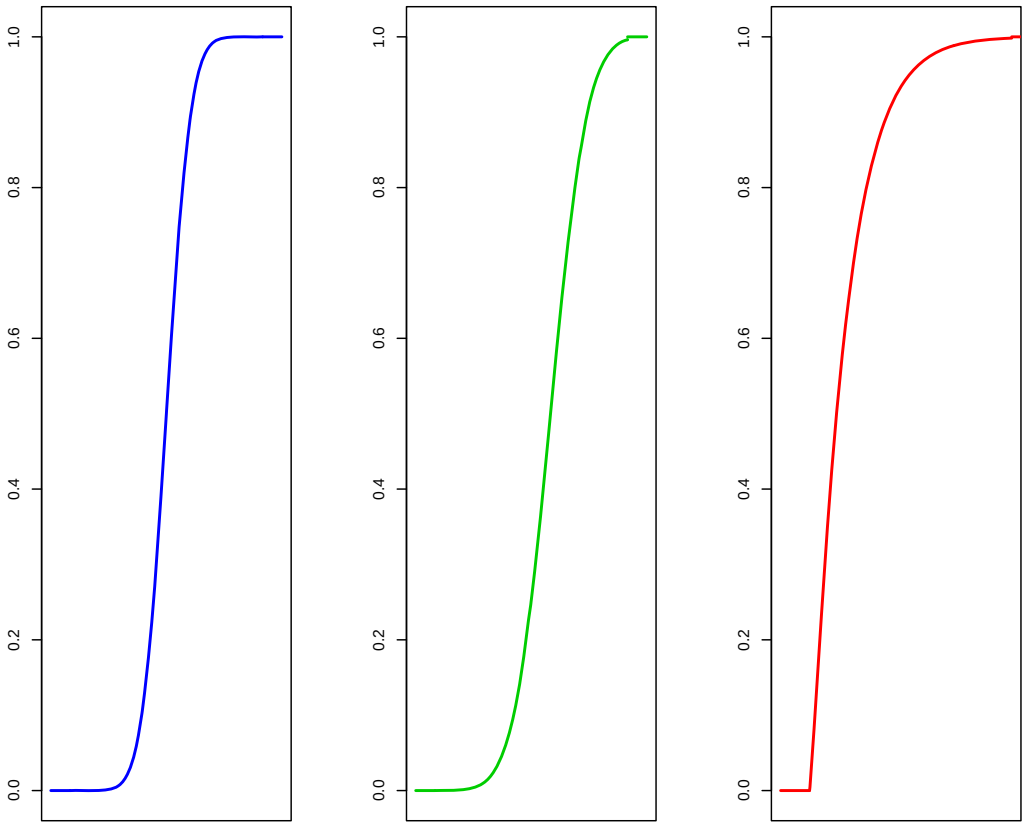
$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}$$

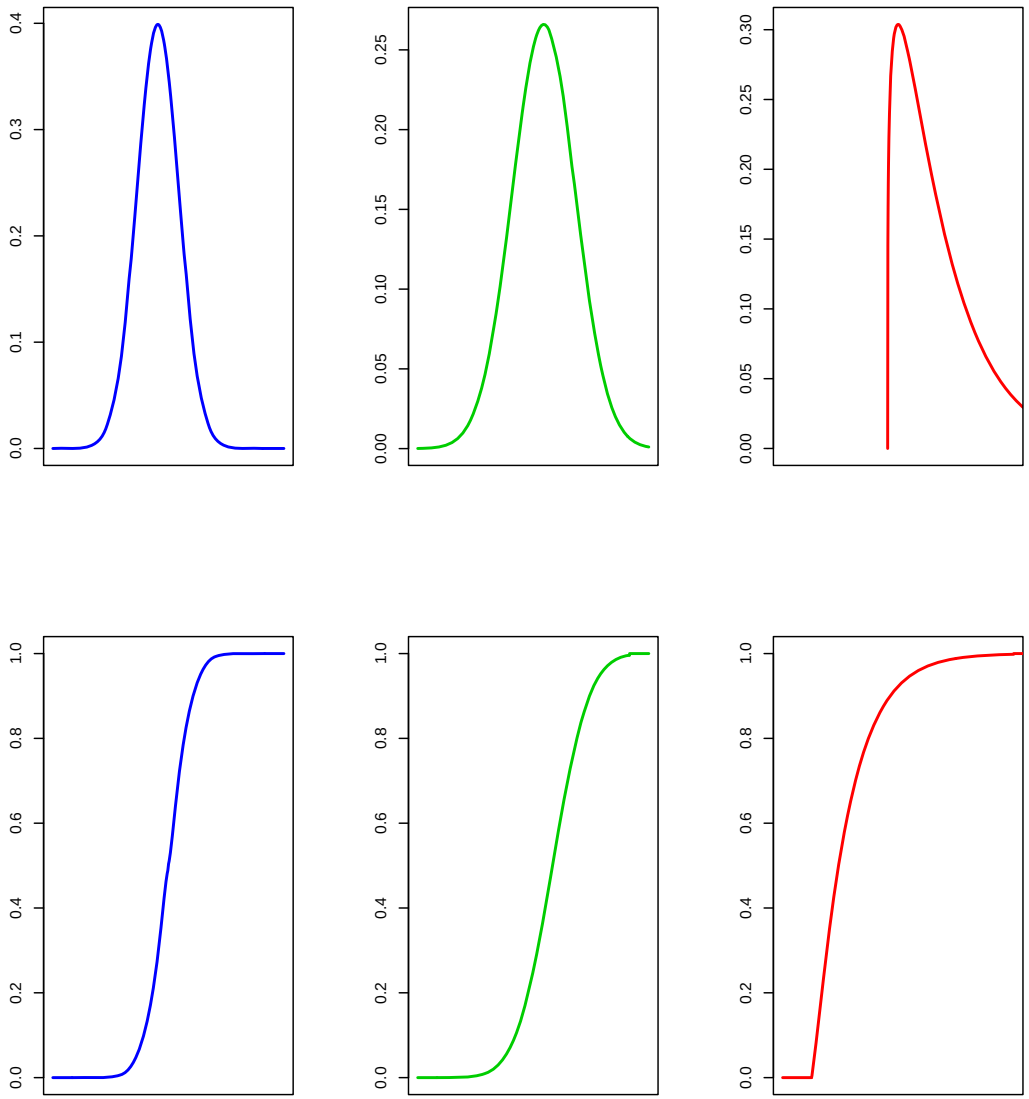
# Distribuciones




Densidades
0.0
0.1
0.2
0.3
0.4

# Distribuciones




0.0
0.1
0.2
0.3
0.4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.2
0.4
0.6
0.8
1.0

$X$ v.a. con distribución $F$:

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}$$

$X$ v.a. con distribución $F$:

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}$$

Propiedades de la distribución:

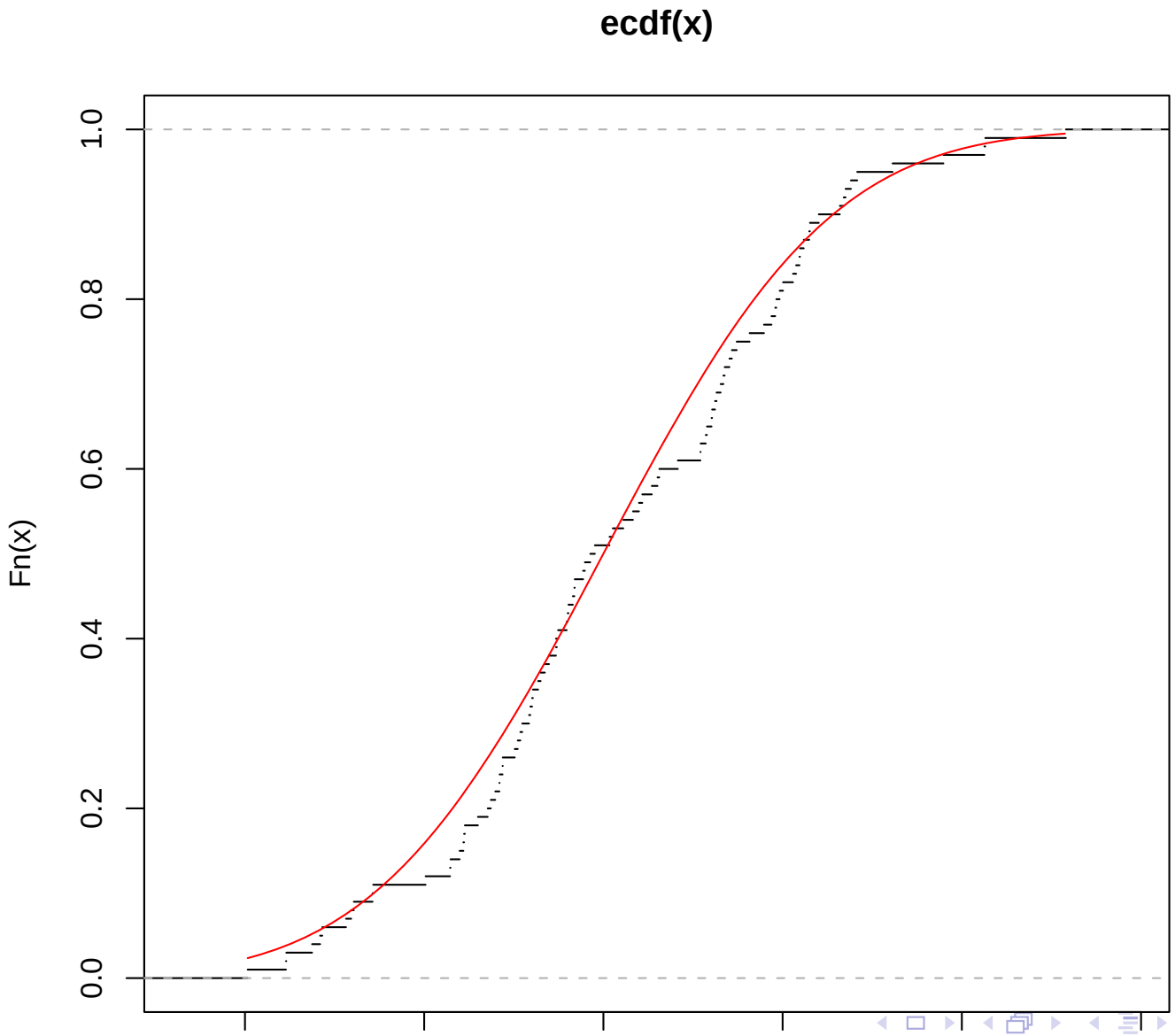
1. $F(x_1) \leq F(x_2)$ para $x_1 < x_2$.
2. $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$.
3. $F$ es continua por la derecha con límite por la izquierda (*cadlag*).

Dada una m.a.s. $X_1, \ldots, X_n$ de $X$, para un $x \in \mathbb{R}$ fijo, queremos estimar $F(x) = \mathbb{P}(X \leq x)$.

Dada una m.a.s. $X_1, \ldots, X_n$ de $X$, para un $x \in \mathbb{R}$ fijo, queremos estimar $F(x) = \mathbb{P}(X \leq x)$.
Denotando por $X_{(1)}, \ldots, X_{(n)}$ la muestra ordenada, se define:

## Función de distribución empírica

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x) = \begin{cases} 0 & \text{si } x \in (-\infty, X_{(1)}), \\ \frac{i}{n} & \text{si } x \in [X_{(i)}, X_{(i+1)}), \\ 1 & \text{si } x \in [X_{(n)}, \infty). \end{cases}$$


ecdf(x)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0

**Ejercicio.** Utilizando la función `ecdf`, calcula la distribución empírica de una muestra de $N(0,1)$.

- Considera distintos tamaños: $n = 10$, $n = 50$, $n = 100$ y compara gráficamente el resultado con la distribución teórica.
- ¿Cuánto vale $F_n$ en el origen?
- ¿Qué ocurre si los datos se generan a partir de una distribución de Cauchy?

Para cada $x$ fijo:

$$\mathbb{E}(F_n(x)) = F(x), \quad \text{Var}(F_n(x)) = \frac{F(x)(1-F(x))}{n}$$

Para cada $x$ fijo:

$$\mathbb{E}(F_n(x)) = F(x), \quad \text{Var}(F_n(x)) = \frac{F(x)(1-F(x))}{n}$$

Además, como $nF_n(x)$ es suma de v.a. independientes con distribución $Bi(n, F(x))$, por el TCL:

$$\sqrt{n}\frac{F_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \xrightarrow{d} N(0,1).$$

Teorema de Glivenko-Cantelli: $F_n(x)$ converge uniformemente a $F(x)$, con probabilidad 1,

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0.$$

Desigualdad de Dvoretzky-Kiefer-Wolfowitz (DKW). Para cualquier $\varepsilon > 0$:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

La desigualdad de DKW puede utilizarse para construir una banda de confianza para $F$. Si igualamos la parte derecha de la desigualdad a $\alpha$:

$$\varepsilon_n^2 = \frac{1}{2n} \log\left(\frac{2}{\alpha}\right)$$

Entonces, la función de distribución $F$ estará comprendida entre $F_n - \varepsilon_n$ y $F_n + \varepsilon_n$ con probabilidad $(1 - \alpha)$. Definimos las bandas:

$$L(x) = \max\{F_n(x) - \varepsilon_n, 0\}, \quad U(x) = \min\{F_n(x) + \varepsilon_n, 1\}$$

por tanto:

$$\mathbb{P}(L(x) \leq F(x) \leq U(x), \forall x) \geq 1 - \alpha.$$

**Ejercicio.** Representa la banda de confianza para $F_n$ a partir de la desigualdad DKW.

```
n<-100
x<-rnorm(n)
Fn<-ecdf(x)
alpha<-0.05
eps<-sqrt(1/(2*n)*log(2/alpha))
t<-seq(min(x),max(x),by=0.01)
L<-pmax(Fn(t)-eps,0)
U<-pmin(Fn(t)+eps,1)
plot(Fn)
points(t,U,type="s")
points(t,L,type="s")
points(t,pnorm(t),t="l",col="red")
```

Por el TCL tenemos que:

$$\sqrt{n}\frac{F_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \sim N(0,1)$$

**Ejercicio.** ¿Cómo construirías bandas de confianza a partir de la distribución límite?

Con la distribución teórica $F$:

```
L2<-pmax(Fn(t)-qnorm(1-alpha/2)*sqrt(pnorm(t)*(1-pnorm(t))/n),0)
U2<-pmin(Fn(t)+qnorm(1-alpha/2)*sqrt(pnorm(t)*(1-pnorm(t))/n),1)
```

Con la distribución empírica $F_n$ (para la varianza):

```
L3<-pmax(Fn(t)-qnorm(1-alpha/2)*sqrt(Fn(t)*(1-Fn(t))/n),0)
U3<-pmin(Fn(t)+qnorm(1-alpha/2)*sqrt(Fn(t)*(1-Fn(t))/n),1)
```

Por el TCL tenemos que:

$$\sqrt{n}\frac{F_n(x)-F(x)}{\sqrt{F(x)(1-F(x))}} \sim N(0,1)$$

$$\mathbb{P}\left(\sqrt{n}\frac{F_n(x)-F(x)}{\sqrt{F(x)(1-F(x))}} \geq t\right) \approx \mathbb{P}(Z \geq t) = 1-\Phi(t)$$

$$= \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-u^2/2}du$$

Por la desigualdad de Tchebychev tenemos que:

$$\mathbb{P}\left(\sqrt{n}\frac{F_n(x)-F(x)}{\sqrt{F(x)(1-F(x))}} \geq t\right) \leq t^{-2}$$

pero se pueden obtener otras tasas más precisas, a partir de la desigualdad anterior.

Si definimos la medida de Dirac como:

$$\delta_x(A) = \begin{cases} 1 & \text{si } x \in A, \\ 0 & \text{si } x \notin A \end{cases}$$

la función de distribución empírica en un punto $x$ se puede escribir como:

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

La integral de una función $\varphi$ con respecto a una medida de dirac $\delta_x$:

$$\int \varphi d\delta_x = \varphi(x).$$

Consideremos la medida (aleatoria):

$$\nu_n = \sum_{i=1}^{n} \frac{1}{n} \delta_{X_i}$$

entonces, la integral con respecto a $\nu_n$ es:

$$\int \varphi d\nu_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i)$$

Si $\varphi(x) = x^k$, se puede obtener el momento de orden $k$ de $X \sim F$ como:

$$\mathbb{E}(X^k) = \int x^k dF(x)$$

Si $\varphi(x) = x^k$, se puede obtener el momento de orden $k$ de $X \sim F$ como:

$$\mathbb{E}(X^k) = \int x^k dF(x)$$

Si substituimos $F$ por la distribución empírica $F_n$ tenemos:

$$\int \varphi(x) dF_n(x) = \int \varphi(x) d\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

que es el estimador empírico.

Si $\varphi_t(x) = \mathbb{I}(x \leq t)$, podemos escribir la función de distribución como:

$$\int \varphi_t(x) dF(x) = F(t)$$

Si $\varphi_t(x) = \mathbb{I}(x \leq t)$, podemos escribir la función de distribución como:

$$\int \varphi_t(x) dF(x) = F(t)$$

Calculando la integral empírica:

$$\int \varphi_t(x) dF_n(x) = F_n(t)$$

Del resultado sobre la distribución de $F_n$ también se obtiene que, para toda función $\varphi$ Borel-medible:

- $\mathbb{E}(\int \varphi dF_n) = \int \varphi dF$
- $\text{Var}(\int \varphi dF_n) = \frac{1}{n}\left(\int \varphi^2 dF - (\int \varphi dF)^2\right) = \sigma^2$
- $\text{Cov}(\int \varphi_1 dF_n, \int \varphi_2 dF_n) = \frac{1}{n}\left(\int \varphi_1\varphi_2 dF - \int \varphi_1 dF \int \varphi_2 dF\right)$
- $\sqrt{n}\left(\int \varphi dF_n - \int \varphi dF\right) \to N(0, \sigma^2)$

**Ejercicio.** Para $s < t$, calcula $\text{Cov}(F_n(s), F_n(t))$.

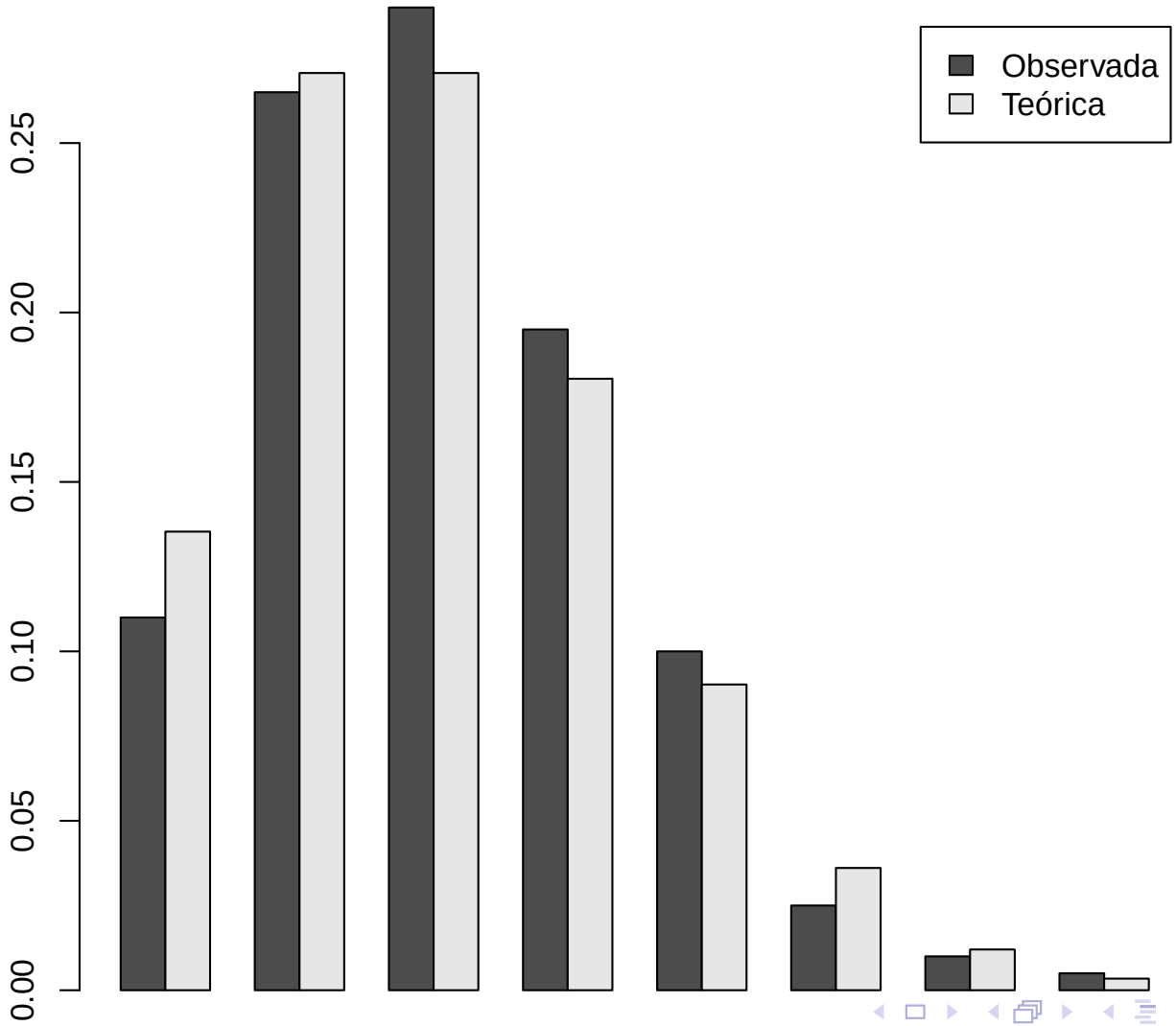
**Ejemplo.** Según estudios llevados a cabo por expertos en defensa, el número de accidentes diarios en un regimiento del ejército en una zona de conflicto sigue una distribución de Poisson de parámetro 2. Se decide comprobar esa hipótesis en un regimiento destinado en Afganistán y se registran los accidentes ocurridos a lo largo de 200 días:

| Nº accidentes | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Nº días | 22 | 53 | 58 | 39 | 20 | 5 | 2 | 1 |

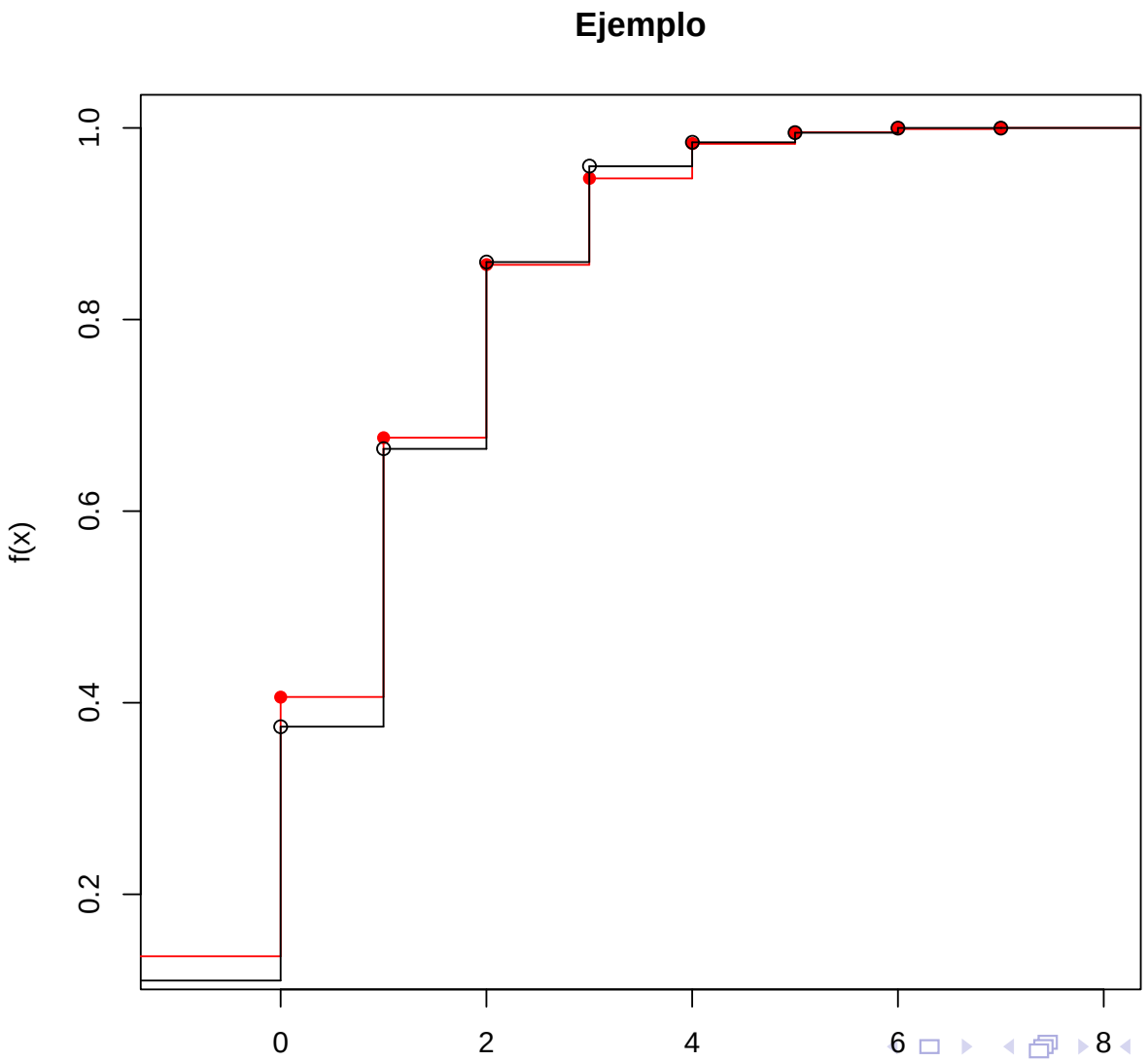
Podemos comprobar gráficamente si la hipótesis es plausible:

Podemos comprobar gráficamente si la hipótesis es plausible:

```
ni<-c(22,53,58,39,20,5,2,1)
pi<-ni/sum(ni)
p0<-dpois(0:7,2)
matriz<-rbind(pi,p0)
colnames(matriz)=0:7
barplot(matriz,beside=TRUE,legend=c("Observada","Teórica"))
```


Observada
Teórica
0.00
0.05
0.10
0.15
0.20
0.25

Podemos comparar la distribución teórica con la empírica:



Planteamos el contraste:

$$\begin{aligned} H_0: &\quad F = Pois(2) \\ H_a: &\quad F \neq Pois(2) \end{aligned}$$

En general: $H_0 : F = F_0$ vs. $H_a : F \neq F_0$, donde $F_0$ es una función perfectamente especificada.

Recordemos el estadístico del teorema de Glivenko-Cantelli, $D_n$:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

$D_n$ es una distancia (medida de discrepancia) entre $F_n$ y $F$. La hipótesis nula se rechazará cuando la diferencia entre la distribución empírica y $F_0$ sea *grande*... y este estadístico es de distribución libre.

- Alternativas unilaterales
- Generalización a dos muestras
- Contraste de normalidad (comparación con Shapiro-Wilks)