Métodos de Remuestreo

Tema 2. Conceptos relacionados con la Distribución Empírica

basado en

B. Efron, R. Tibshirani (1993). An Introduction to the bootstrap.

O. Kirchkamp (2017). Resampling methods.

Curso 2017/18

Parámetros, distribuciones y el principio de plug-in

- El mejor modo de visualizar una muestra aleatoria es en términos de una **población finita**: un universo U de unidades individuales $U_1, U_2, \ldots, U_N$ cada una de las cuales tiene la misma probabilidad de ser seleccionada.
- En cada unidad U_i se mide una variable de interés X_i de modo que se obtiene un *censo* $X_1, X_2, \ldots, X_N$, es decir, $\mathbb{X}$.
- Una muestra aleatoria de tamaño n es una colección de n unidades $u_1, u_2, \ldots, u_n$ seleccionadas al azar del universo U.
- En cada unidad seleccionada u_i se toma una medida de interés x_i de modo que una muestra se denota como $\mathbf{x}$.
- Los problemas en estadística en general se refieren a estimar algún aspecto de la distribución de probabilidad F en base a una muestra.

Ejemplo sobre el principio de *plug-in*

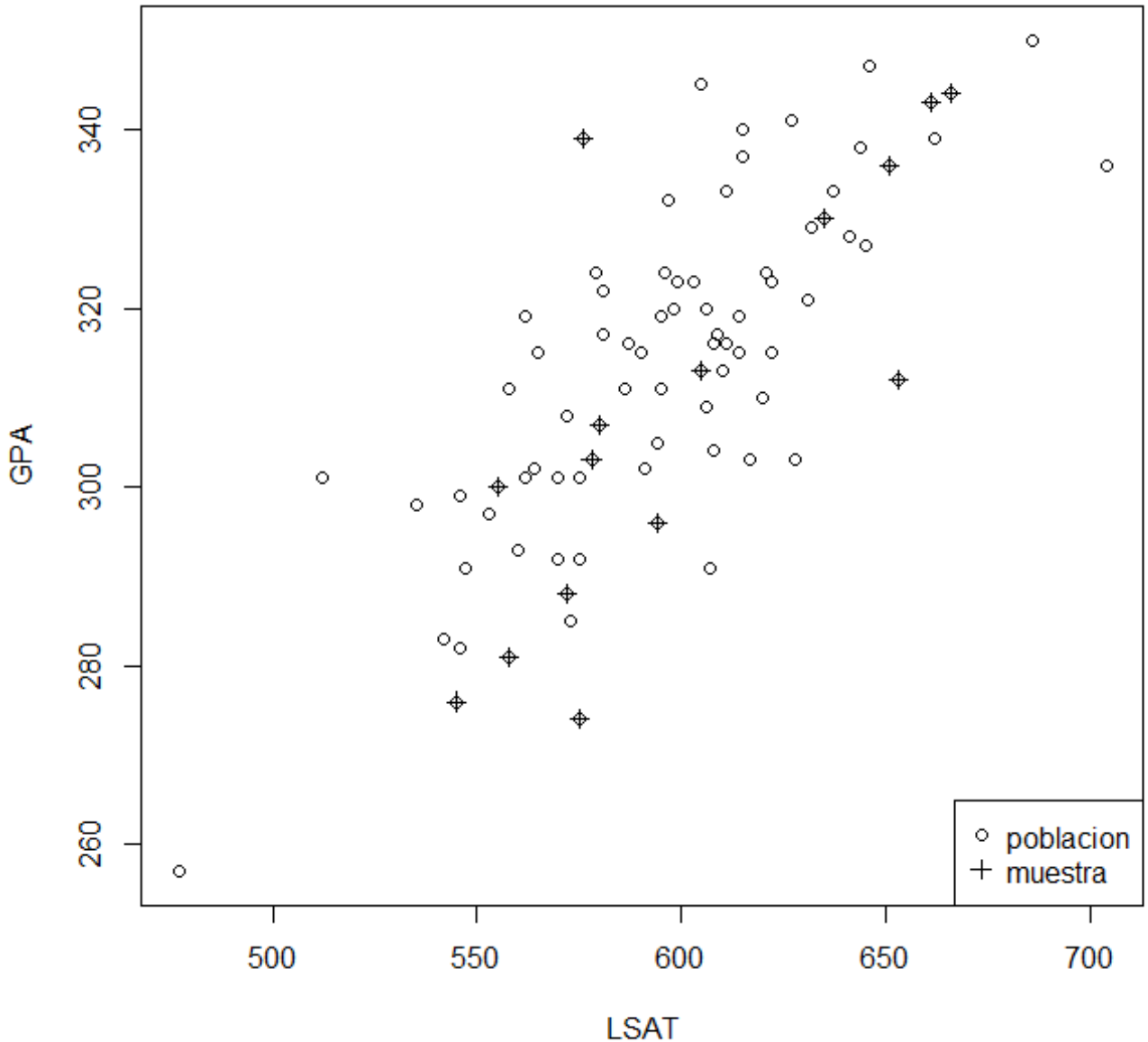
- Se toma el ejemplo de las universidades con máster en leyes que está incluido en el libro de Efron y Tibshirani.

- Se trata de una población compuesta por 82 universidades en relación un máster en *Leyes* (está en la librería `bootstrap` de R con el nombre `law82`), donde `GPA` es la puntuación media en los cursos de grado, y `LSAT` es la calificación de admisión.

- La muestra contiene 15 observaciones.

```
library(bootstrap)

with(law82, plot(100*GPA ~ LSAT, ylab="GPA"))
with(law, points(100*GPA ~ LSAT, pch=3))
legend("bottomright", c("poblacion", "muestra"),
pch=c(1,3))
```


GPA
260
280
300
320
340
500
550
600
650
700
LSAT
poblacion
muestra

Ejemplo sobre el principio de *plug-in*

- Interesa calcular la correlación entre GPA (la puntuación media en los cursos de grado) y LSAT (calificación de admisión).

- Con la verdadera puntuación poblacional:

```
with(law82, cor(GPA,LSAT))
```

```
[1] 0.7599979
```

- El estimador *plug-in* es

```
with(law, cor(GPA,LSAT))
```

```
[1] 0.7763745
```

Función de distribución empírica

- La distribución empírica denominada $\widehat{F}$ es un estimador simple de la función de distribución teórica F.
- El principio de *plug-in* consiste en estimar algún aspecto de F como la media, mediana etc. mediante $\widehat{F}$.
- El bootstrap es una aplicación directa de este principio.
- Supongamos que se observa una muestra aleatoria de tamaño n con función de distribución F

$$F \rightarrow (x_1, x_2, \ldots, x_n)$$

Función de distribución empírica

- La función de distribución empírica $\widehat{F}$ se define como la distribución discreta que asigna probabilidad $\frac{1}{n}$ a cada valor x_i tal que $i = 1, 2, \ldots, n$
- De este modo $\widehat{F}$ asigna a un conjunto A del espacio muestral de x la probabilidad empírica

$$\widehat{P}(A) = \frac{\#\{x_i \in A\}}{n}$$

- Esa es la proporción de la muestra observada **x** que ocurre en A.

Función de distribución empírica

```
# Simulo datos de calificaciones

mu = 6.5
sigma = 0.5

y = rnorm(n=20, mean=mu, sd=sigma)
y = round(y,3)
t = mean(y)
```

```
cat("La muestra ordenada es", sort(y),
"\n y se obtiene una media muestral igual a ",
t, "\n")
```

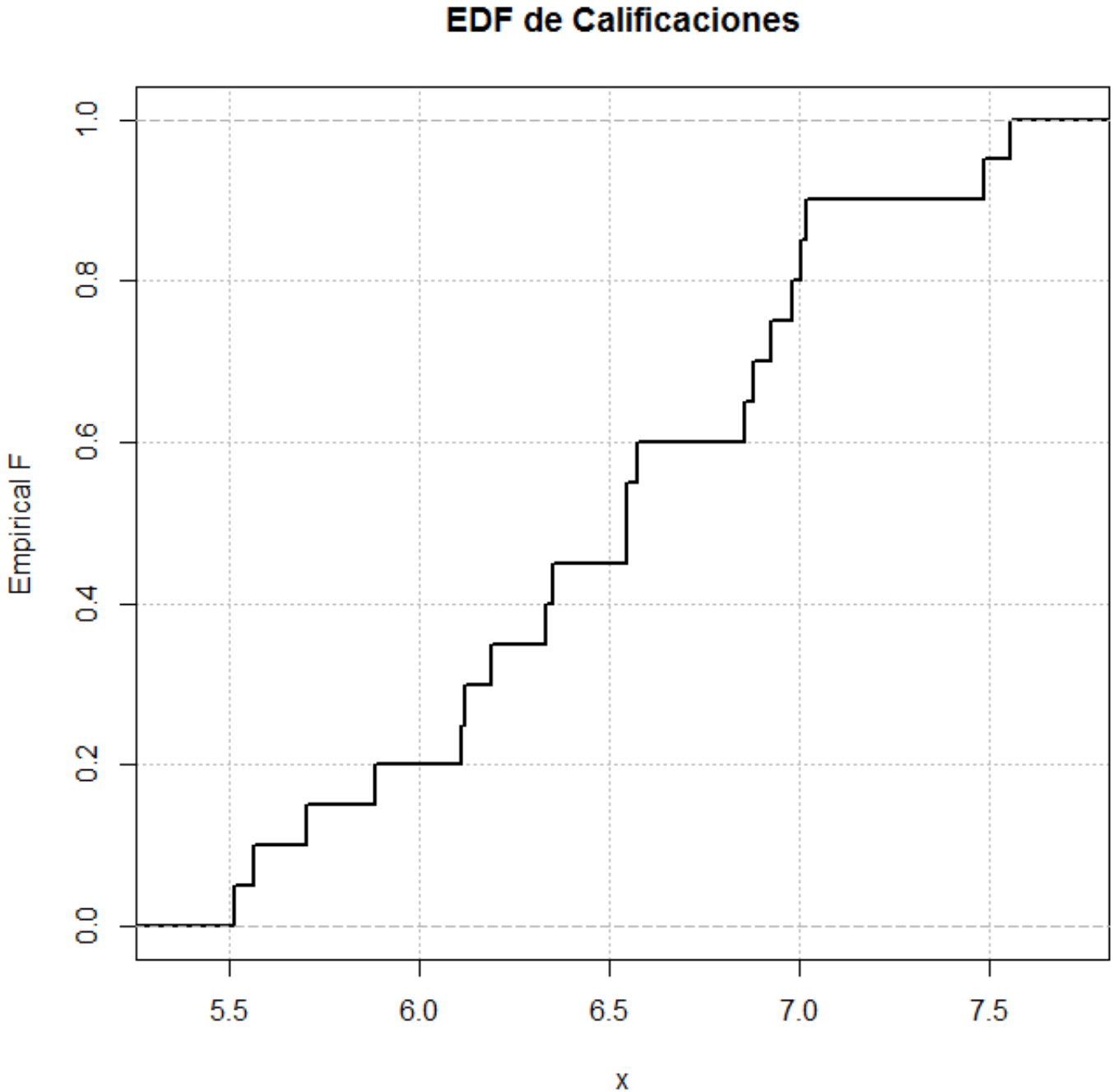
```
La muestra ordenada es 5.72 5.738 5.948 6.163 6.171 6.357 6.498
    6.576 6.607 6.613 6.615 6.686 6.7 6.749 6.847 6.908 6.926
    7.2 7.22 7.306
y se obtiene una media muestral igual a  6.5774
```

Función de distribución empírica

Se puede dibujar la correspondiente función de distribución empírica

```
# EDF

plot.ecdf(x=y, verticals=TRUE, do.p=FALSE,
main="EDF de Calificaciones", lwd=2,
panel.first=grid(col="gray",lty="dotted"),
ylab="Empirical F")
```

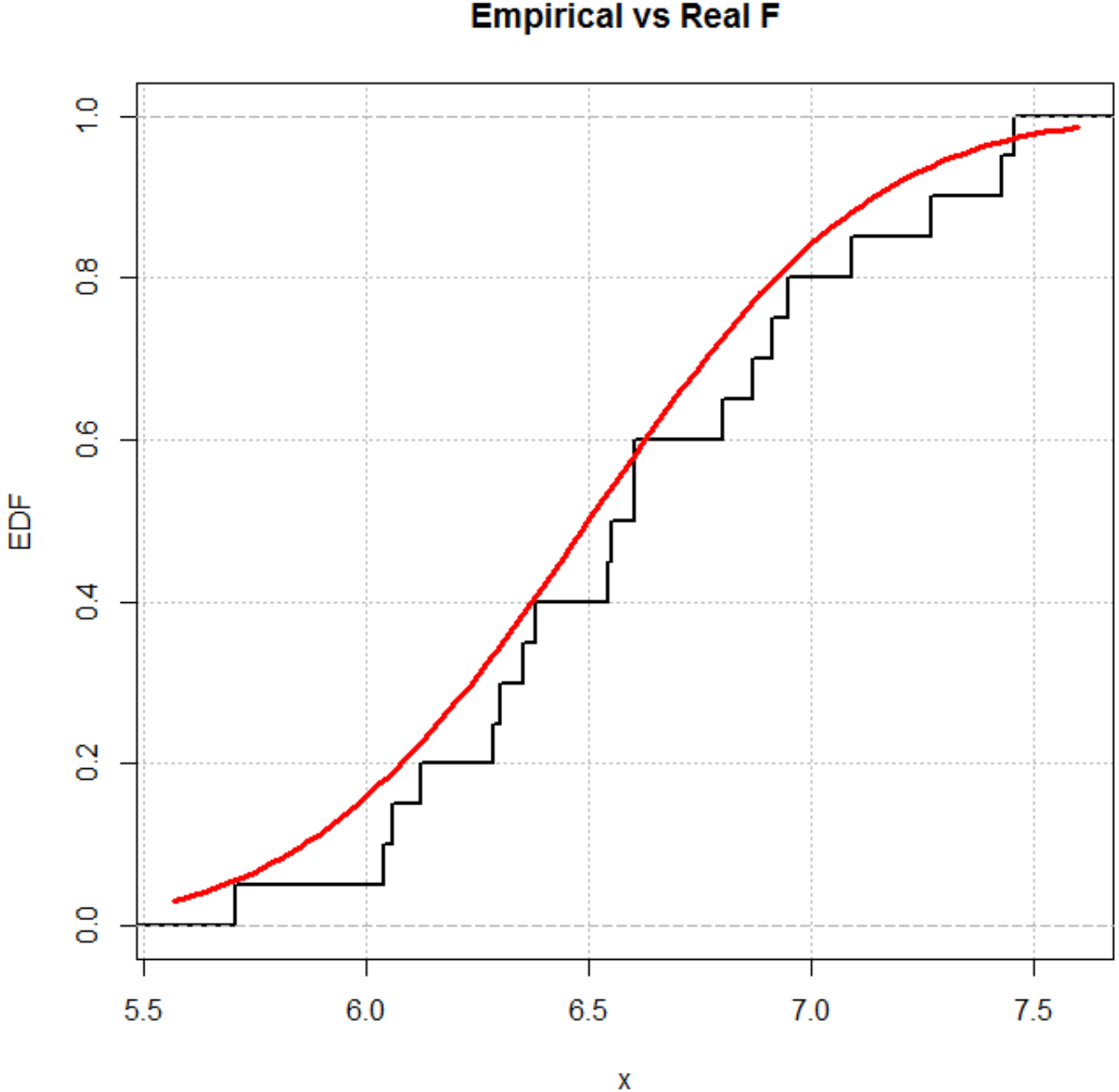

EDF de Calificaciones
Empirical F
0.0
0.2
0.4
0.6
0.8
1.0
5.5
6.0
6.5
7.0
7.5
x

Función de distribución empírica

Se puede dibujar la correspondiente función de distribución empírica junto con la curva de la función de distribución real.

```
plot.ecdf(x=y, verticals=TRUE, do.p=FALSE,
main="Empirical vs Real F", lwd=2, xlab="x",
panel.first = grid(nx=NULL, ny=NULL,
col="gray", lty="dotted"), ylab="EDF")

curve(expr=pnorm(x, mean=mu, sd=sigma), col="red",
add=TRUE, lw=3)
```


Empirical vs Real F
EDF
0.0
0.2
0.4
0.6
0.8
1.0
5.5
6.0
6.5
7.0
7.5
x

Función de distribución empírica

- Se define la función de distribución empírica como

$$\widehat{F}_n(x) = \frac{\text{Número de elementos de la muestra} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}\{x_i \leq x\}$$

donde $\mathbf{I}\{A\}$ es la función indicatriz del suceso A.

- En general, $\widehat{F}_n(x)$ se puede considerar como una función de distribución discreta que asigna probabilidad igual a $\frac{1}{n}$ a cada uno de los n valores $x_1, \ldots, x_n$.

- Así $\widehat{F}_n(x)$ es una función escalón con un salto de tamaño $\frac{1}{n}$ en cada punto x_i para $i = 1, \ldots, n$.

Función de distribución empírica

- Si se ordenan los valores de la muestra de menor a mayor $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ entonces then $\widehat{F}_n(x) = 0$ para $x < x_{(1)}$
- $\widehat{F}_n(x)$ salta al valor $1/n$ en $x = x_{(1)}$ y se mantiene igual a $1/n$ para $x_{(1)} \leq x < x_{(2)}$
- $\widehat{F}_n(x)$ salta al valor $2/n$ en $x = x_{(2)}$ y se mantiene igual a $2/n$ para $x_{(2)} \leq x < x_{(3)}$ y así sucesivamente
- Si se fija el valor de x entonces la variable aleatoria $\mathbf{I}\{x_i \leq x\}$ es una v.a. Bernoulli de parámetro $p = F(x)$
- Entonces $n\widehat{F}_n(x)$ es una v.a **binomial** de media $nF(x)$ y varianza $nF(x)(1 - F(x))$.

Propiedades de la función de distribución empírica

- Así,

$$E\left[n\widehat{F}(x)\right] = nF(x) \Rightarrow \boxed{E\left[\widehat{F}(x)\right] = F(x)}$$

$$Var\left[n\widehat{F}(x)\right] = nF(x)(1-F(x)) \Rightarrow \boxed{Var\left[\widehat{F}(x)\right] = \frac{1}{n}F(x)(1-F(x))}$$

  de modo que $\widehat{F}(x)$ es un estimador insesgado de $F(x)$.

- Si se denota como $F(x)$ la función de distribución de la v.a. de la que procede la muestra entonces, para todo número $(-\infty < x < \infty)$, la probabilidad de que una observación dada X_i sea menor o igual que x es $F(x)$.
- Por tanto, por la *ley de los grandes números*, cuando $n \to \infty$, la proporción $\widehat{F}_n(x)$ de observaciones en la muestra que son menores o iguales que x convergen en probabilidad a $F(x)$.

$$\widehat{F}_n(x) \xrightarrow{p} F(x) \qquad -\infty < x < \infty$$

Propiedades de la función de distribución empírica

- Pero se tiene un resultado más potente: $\widehat{F}_n(x)$ converge a $F(x)$ de manera ***uniforme*** para todos los valores de x
- **Lema de Glivenko-Cantelli** Sea $\widehat{F}_n(x)$ una función de distribución empírica de una m.a.s $X_1, \ldots, X_n$ de una función de distribución F entonces,

$$\boxed{D_n = \underset{-\infty<x<\infty}{\text{Sup}} \left|\widehat{F}_n(x) - F(x)\right| \xrightarrow{p} 0}$$

- **Nota**: Antes de que los valores $X_1, \ldots, X_n$ hayan sido observados D_n es una v.a.
- Cuando el tamaño muestral n es grande la función de distribución empírica $\widehat{F}_n(x)$ está muy próxima a $F(x)$ sobre la recta real.
- Así, cuando se desconoce la función de distribución $F(x)$ se puede considerar que $\widehat{F}_n(x)$ es un estimador muy eficiente de $F(x)$.

Simulaciones de la función de distribución empírica

- Se define una función para calcular la función de distribución empírica en cada punto:

```
x = rpois(20,3) # ej. tomas una m.a.s de una Poisson
P = ecdf(x)
P(3)
```

```
[1] 0.5
```

```
acumula.dist = function(muestra, z){
    cuento = 0
    for(t in muestra){ if(t<=z) cuento = cuento+1 }
    return(cuento/length(muestra))
}
acumula.dist(x, 3)
```

- Para simular de la función de distribución empírica una vez observado vector x, se puede usar la función `sample`.

```
sample(x, size=20, replace=TRUE)
```

Propiedades de la función de distribución empírica

- Supongamos que se tiene un conjunto de observaciones $X_1, \ldots, X_n$ procedente de una m.a.s. de una población con función de distribución F.

- Dado un un estadístico de interés $T_n = T_n(X_1, \ldots, X_n)$, se trata de estimar la distribución de una función de F y T_n, digamos $R_n(T_n, F)$:

$$H_n(x) = P\{R_n \leq x\},$$

- Por ejemplo, R_n podría ser la cantidad pivotal que se usa para construir los intervalos para la media asumiendo normalidad:

$$R_n = \frac{\overline{X}_n - \mu}{\sigma_0/\sqrt{n}}$$

Propiedades de la función de distribución empírica

- Entonces, se obtienen muestras bootstrap $X_1^*, \ldots, X_n^*$ de la distribución empírica basada en $X_1, \ldots, X_n$ de modo que se puede definir $T_n^* = T_n(X_1^*, \ldots, X_n^*)$ y $R_n^* = R_n(T_n^*, F_n)$, donde F_n es la función de distribución empírica.

- El estimador bootstrap de H_n se calcula como

$$\widehat{H}_n(x) = P_* \{R_n^* \leq x\},$$

donde P_* es la distribución basada en las muestras bootstrap.

- Se observa que $\widehat{H}_n(x)$ depende de la distribución empírica F_n. Así, $\widehat{H}_n(x)$ cambia cuando los datos $\{x_1, \ldots, x_n\}$ varían.

Propiedades de la función de distribución empírica

- Supongamos que se trata de estimar la varianza de un estimador $\widehat{\theta} = \theta(x_1, \ldots, x_n)$.

- La varianza teórica viene dada por

$$Var(\widehat{\theta}) = \int \cdots \int \left\{\theta(x_1, \ldots, x_n) - E(\widehat{\theta})\right\}^2 dF(x_1) \cdots dF(x_n)$$

donde

$$E(\widehat{\theta}) = \int \cdots \int \theta(x_1, \ldots, x_n) dF(x_1) \cdots dF(x_n)$$

Propiedades de la función de distribución empírica

- La solución natural es usar como estimador *plug-in* la distribución empírica $\widehat{F}$

$$\widehat{Var}(\widehat{\theta}) = \int \cdots \int \left\{\theta(x_1, \ldots, x_n) - \widehat{E}(\widehat{\theta})\right\}^2 d\widehat{F}(x_1) \cdots d\widehat{F}(x_n)$$

- es decir, al ser $\widehat{F}$ una distribución discreta

$$\widehat{Var}(\widehat{\theta}) = \frac{1}{n^n} \sum_j \left\{\theta(x_1^j, \ldots, x_n^j) - \widehat{E}(\widehat{\theta})\right\}^2$$

donde $(x_1^j, \ldots, x_n^j)$ varía entre todas las posibles combinaciones de los datos muestrales.

Propiedades de la función de distribución empírica

- A no ser que n sea pequeño, los cálculos anteriores pueden llevar bastante tiempo computacional.
- Sin embargo, se pueden aproximar las expresiones mediante integración Monte Carlo.
- Se aproxima la integral tomando muestras aleatorias de tamaño n de la función $\widehat{F}$ y calculando la media muestral del integrando.
- Por la *Ley de los Grandes Números* esta aproximación converge al verdadero valor de la integral cuando $n \to \infty$.

Propiedades de la función de distribución empírica

- ¿A qué se parece una muestra aleatoria tomada de $\widehat{F}$?
- Como $\widehat{F}$ asigna igual masa de probabilidad a cada valor observado x_i, el hecho de tomar una muestra aleatoria de $\widehat{F}$ es equivalente a tomar n valores con reemplazamiento de $x_1, \ldots, x_n$.
- De hecho, a la técnica de tomar nuevas muestras a partir de la muestra original es a lo que realmente se denomina ***remuestreo***.

Intervalos de confianza basados en la función de distribución empírica

- ***Teorema de Dvoretzky-Kiefer-Wolfowitz*** (DKW):

  Sea $X_1, \ldots, X_n$ una muestra aleatoria de una v.a. con función de distribución F. Entonces, para todo $\varepsilon > 0$

$$\boxed{P\left(\sup_x \left|F(x) - \widehat{F}_n(x)\right| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}}$$

- Se pueden construir así los intervalos de confianza para $\widehat{F}$.

Intervalos de confianza basados en la función de distribución empírica

- Se define

$$L(x) = \text{máx}\left\{\widehat{F}_n(x) - \varepsilon_n, 0\right\}$$

$$U(x) = \text{mín}\left\{\widehat{F}_n(x) + \varepsilon_n, 1\right\}$$

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

- de este modo, para toda función de distribución F y para todo x se tiene que

$$\boxed{P\left(L(x) \leq F(x) \leq U(x)\right) \geq 1 - \alpha}$$

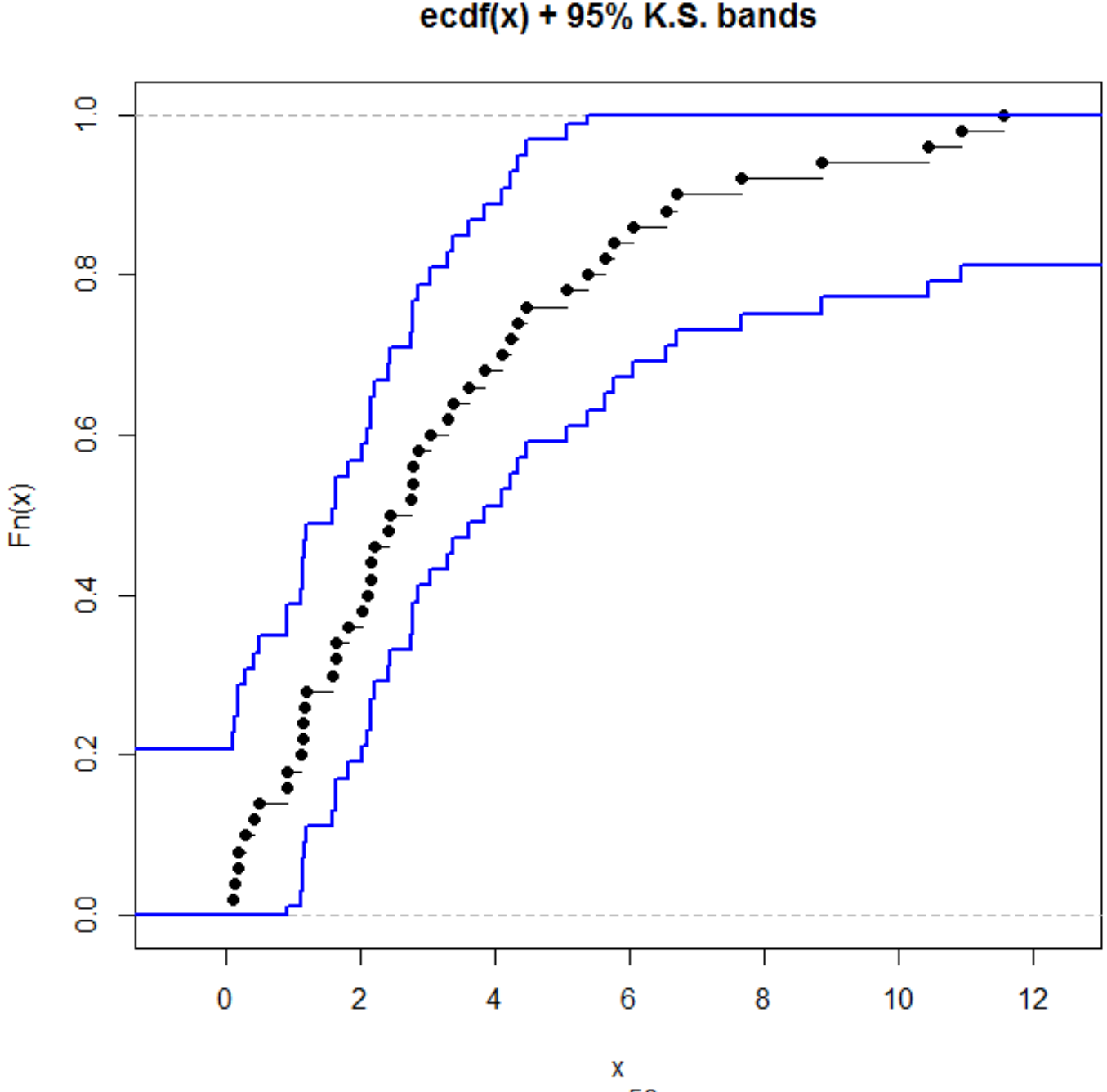
Intervalos de confianza basados en la función de distribución empírica

```
library(sfsmisc)

# Simulas datos de una v.a. chi cuadrado con 3 g.l.
x = rchisq(50,3)

X11()
ecdf.ksCI(x, ci.col="blue", lwd=2)
```

Se pueden programar fácilmente las cotas, pero los intervalos son amplios para tamaños muestrales pequeños.


ecdf(x) + 95% K.S. bands
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
x
n = 50

Intervalos de confianza basados en la función de distribución empírica

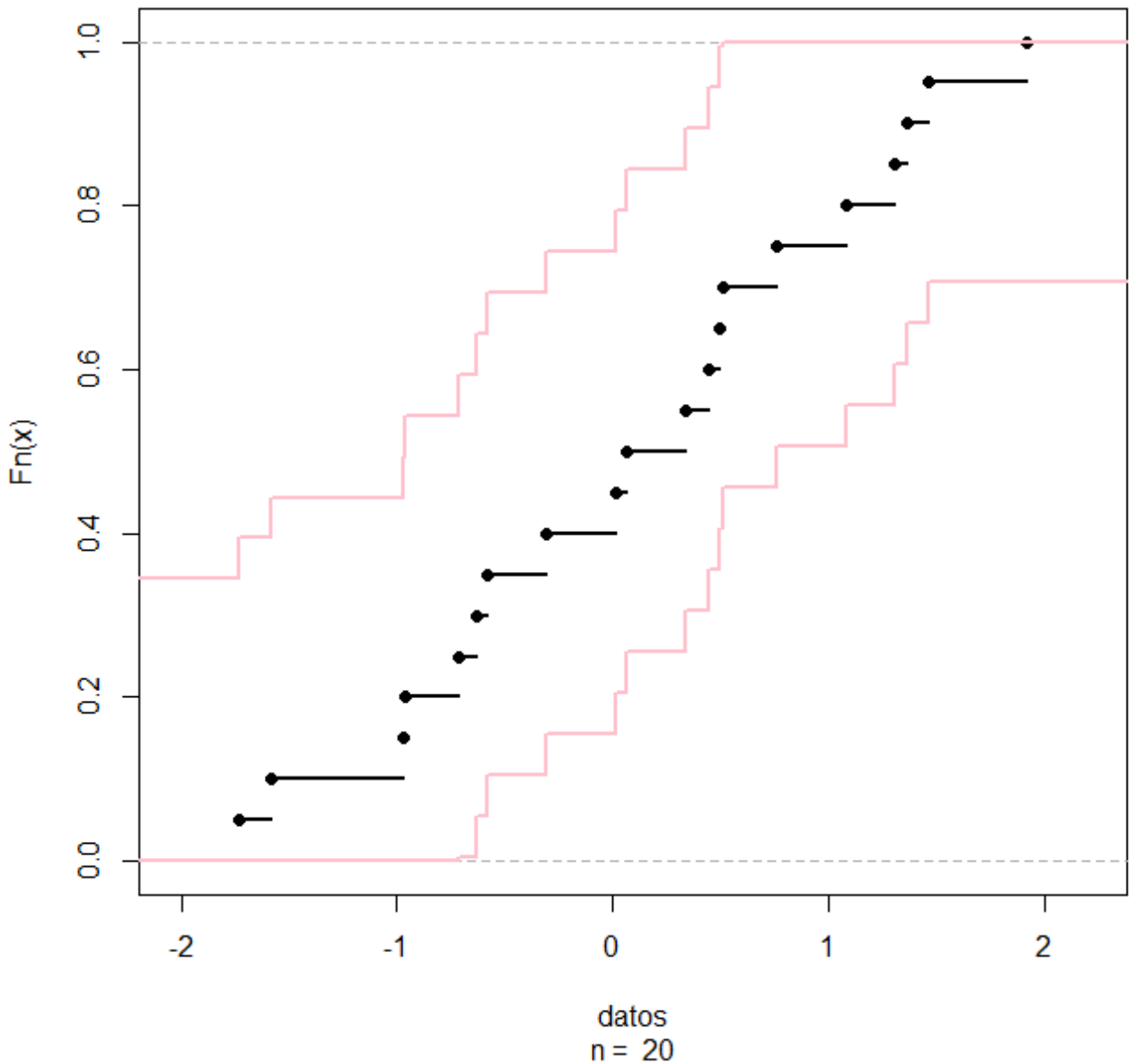
```
dkw_cota = function(datos, x, alfa){
    P = ecdf(datos)
    F_boina = P(x)
    epsilon = sqrt(log(2/alfa)/(2*length(datos)))
    inf_cota = pmax(F_boina - epsilon, 0)
    sup_cota = pmin(F_boina + epsilon, 1)
    return(c(inf_cota, sup_cota))
}

datos = rt(20,3)      # Simulas de una t de Student

dkw_cota(datos, -0.5, 0.05)
```

```
0.04631927 0.65368073
```

```
ecdf.ksCI(datos, ci.col="pink", lwd=2)
```


ecdf(datos) + 95% K.S. bands
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
datos
n = 20