Ejercicios 4

Test de Kolmogorov-Smirnov

Joaquin Cavieres G.

Introducción

El test de Kolgomorov-Smirnov (K-S) generalmente es utilizado para determinar si una muestra aleatoria proviene desde una población en particular con una función de distribución específica. El test K-S estba basado en la función de distribución empírica (ECDF por sus siglas en inglés) para un conjunto determinado de observaciones.

Considere n puntos ordenadas de datos observados $X_1, ..., X_n$. Dado lo anterior es que la ECD, que vamos a denotar como $F_n(x)$, puede calcularse como:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x}, \tag{1}$$

que es un estimador no paramétrico de F.

Adicionalmente a este estimador de F, existe una medida de proximidad entre $F_n(x)$ y F (la cual está especificada en H_0 generalmente como F_0), la cual nos indicaria sus relaciones en terminos de "distancia". Si esta "distancia" estimada entre F_n y F_0 es grande , entonces existirían evidencias de rechazar H_0 ya que damos por verdadero que F_0 es igual a una cierta función de distribución de referencia.

Caracteristicas del test K-S

Una de las principales caracteristicas de este test que la distribución del estadístico k-S no depende de la funci´pn de distribución acumulada que se esta contrastando. Además, es una prueba exacta, en comparación con la prueba X^2 que depende de un tamaño de muestra adecuado para que sus aproximaciones sean válidas.

Este test también posee algunas limitantes, como por ejemplo, sólo puede ser aplicado a funciones continuas o que es sensible en el centro de la distribución que en las colas. Sin embargo una de sus principales limitantes es que la distribución debe especificarse completamente. Esto quiere decir que si los parámetros de ubicación, escala y forma son estimados a partir de los datos observados, la región crítica del test K-S ya no es válida. Generalmente debe estimarse mediante simulación.

Como observación, esta prueba si bien ha sido desarrollada para funciones continuas, también se ha generalizado a distribuciones discretas con datos censurados y agrupados, pero en esta clase no veremos esta caracteristica.

Definición del test Kolgomorov-Smirnov (K-S)

El test K-S podemos definirlo de la siguiente manera. Primero debemos espeficiar una hipótesis nula H_0 en donde se propone que los datos siguien una función de distribución específica. Segundo debemos especificar una hipótesis alternativa H_1 la cual indica que los datos observados no provienen de la función de distribución específicada en H_0. Y por último, el estadístico de prueba K-S esta definido como:

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

Si proponemos que $H_0 : F = F_0$ y se cumple, entonces D_n tiende a ser pequeño, al contrario, si $H_0 : F \neq F_0$ entonces se espera que D_n sea grande lo que nos llevaría a rechazar H_0.

El calculo de D_n se obtiene al determinar la diferencia máxima entre F_0 y F_n cuando $x = X$ para cierto X_i. En primera instancia debemos ordenar la muestra y considerar:

$$\begin{aligned} D_n &= \text{máx}(D_n^+, D_n^-), \\ D_n^+ &:= \sqrt{n} \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - U_{(i)} \right\}, \\ D_n^- &:= \sqrt{n} \max_{1 \leq i \leq n} \left\{ U_{(i)} - \frac{i-1}{n} \right\}, \end{aligned} \tag{6.2}$$

Si la distribución continua bajo H_0 se cumple en función de la propuesta en F_0, entonces D_n tiene una CDF asintótica dada por una función K de kolgomorov-Smirnov. Eso es:

$$\lim_{n \to \infty} \mathbb{P}[D_n \leq x] = K(x) := 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 x^2}. \tag{6.3}$$

Como se comentó previamente, el estadístico K-S es un test de distribución libre (*distribution-free* en inglés) por que la distribución bajo H_0 no depende de F_0 pero sólo si F_0 es una función continua y la muestra aleatoria $X_1, ..., X_n$ también es continua. Si estos supuestos son cumplidos entonces la muestra $X_1, ..., X_n \overset{H_0}{\sim} F_0$ genera una muestra i.i.d $U_1, \ldots, U_n \overset{H_0}{\sim} \mathcal{U}(0, 1)$. Como consecuencia de esto la distribución de D_n no depende de F_0.Si F_0 no es continua entonces la función K no es la verdadera distribución asintótica.

Ejercicio 1

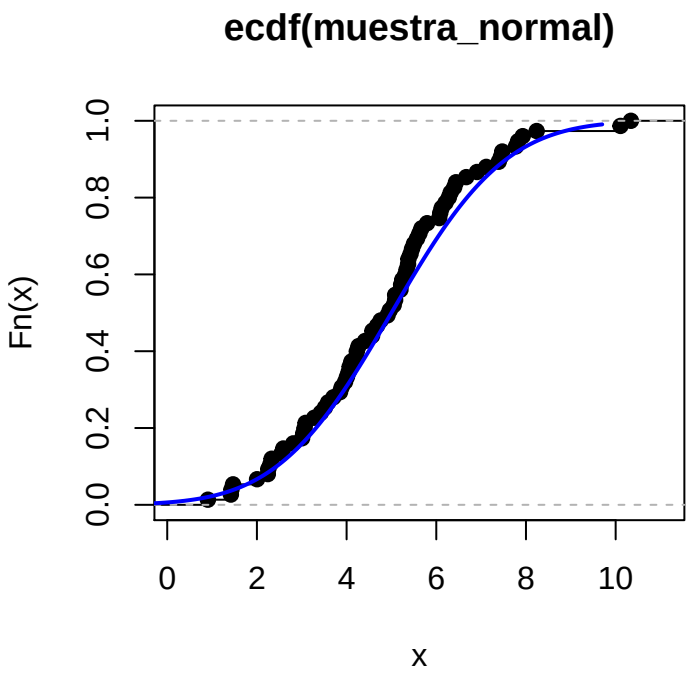
```r
# Creamos una muestra aleatoria con distribución Normal~(0,1)
set.seed(1000)
muestra_normal <- rnorm(75,mean=5, sd=2)

F75 = ecdf(muestra_normal)
plot(F75)
knots(F75)
```

```
##  [1]  0.9068292  1.4193541  1.4323117  1.4676028  2.0033791  2.2437949
##  [7]  2.2537645  2.3032819  2.3279179  2.5459680  2.5882869  2.8106126
## [13]  3.0073960  3.0351443  3.0636342  3.0861447  3.2783245  3.4268913
## [19]  3.5167571  3.5871266  3.7063501  3.8530050  3.8910226  3.9653871
## [25]  4.0085062  4.0482642  4.0696981  4.1084435  4.2153239  4.2290214
## [31]  4.2676386  4.4113176  4.5681732  4.5770928  4.6813128  4.7582554
## [37]  4.9126171  4.9629888  5.0498637  5.0822526  5.0824742  5.2011560
## [43]  5.2152473  5.2427624  5.3101574  5.3401150  5.3709301  5.3785772
## [49]  5.4263082  5.4593333  5.5034228  5.5709152  5.6234024  5.6698843
## [55]  5.7952857  6.0651435  6.0878569  6.1219514  6.2063225  6.2787768
## [61]  6.3219183  6.3988591  6.4395014  6.6684947  6.9073606  7.1152024
## [67]  7.3921729  7.4418713  7.4708838  7.7777326  7.8234714  7.9275507
## [73]  8.2384175 10.1079760 10.3401433
```

```r
# Vemos los valores de x ordenados
sort(muestra_normal)
```

```
##  [1]  0.9068292  1.4193541  1.4323117  1.4676028  2.0033791  2.2437949
##  [7]  2.2537645  2.3032819  2.3279179  2.5459680  2.5882869  2.8106126
## [13]  3.0073960  3.0351443  3.0636342  3.0861447  3.2783245  3.4268913
## [19]  3.5167571  3.5871266  3.7063501  3.8530050  3.8910226  3.9653871
## [25]  4.0085062  4.0482642  4.0696981  4.1084435  4.2153239  4.2290214
## [31]  4.2676386  4.4113176  4.5681732  4.5770928  4.6813128  4.7582554
## [37]  4.9126171  4.9629888  5.0498637  5.0822526  5.0824742  5.2011560
## [43]  5.2152473  5.2427624  5.3101574  5.3401150  5.3709301  5.3785772
## [49]  5.4263082  5.4593333  5.5034228  5.5709152  5.6234024  5.6698843
## [55]  5.7952857  6.0651435  6.0878569  6.1219514  6.2063225  6.2787768
## [61]  6.3219183  6.3988591  6.4395014  6.6684947  6.9073606  7.1152024
## [67]  7.3921729  7.4418713  7.4708838  7.7777326  7.8234714  7.9275507
## [73]  8.2384175 10.1079760 10.3401433
```

```r
# Hacemos un gráfico de la ecdf
points = seq(-0.5,9.7,0.01)
norm_cdf = pnorm(points,mean=5,sd=2)
lines(points, norm_cdf, col= "blue", lwd = 2)
```




```r
# Distribución exponencial
muestra_exp <- rexp(100, rate = 1)
ks.test(muestra_exp,"pnorm", mean=0, sd=2)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  muestra_exp
## D = 0.50126, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
# ¿Rechazamos H0 si consideramos un alpha de 0.05?


# Probamos con otra hipotesis en base a "pexp"
ks.test(muestra_exp,"pexp", rate = 1)
```

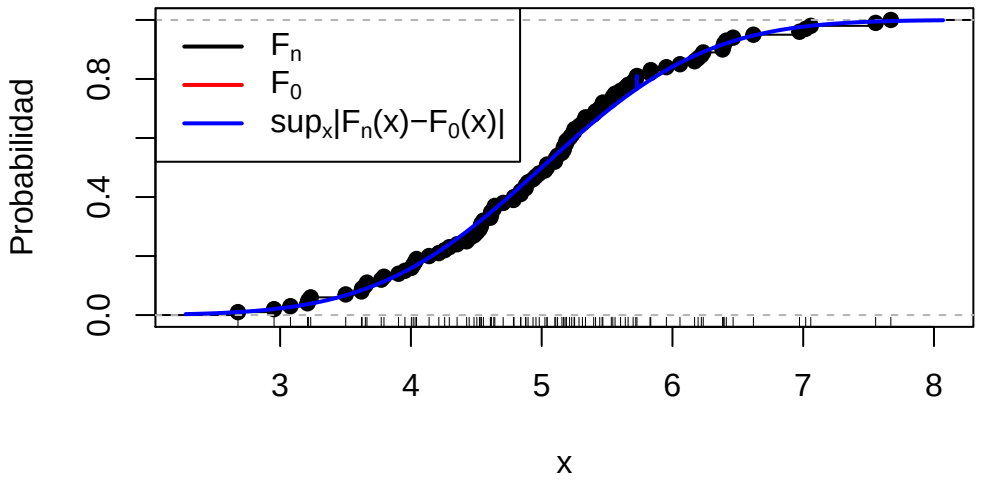
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  muestra_exp
## D = 0.13521, p-value = 0.05167
## alternative hypothesis: two-sided
```

```r
# p-value = 0.05167
# No rechazamos H0 ya que 0.05167 > a 0.05 (alpha)
```

Ejercicio 2

Vamos a contrastar la hipotesis nula de $H_0 : F = F_0$ vs $H_1 : F \neq F_0$. En este caso vamos simular una variable aleatoria con distribución Normal con media 5 y desviación estandar de 1.

```
# Distribución Normal
set.seed(1000)            # Para que sea reproducible el ejercicio
n = 100
mu = 5
sd = 1
sample = rnorm(n, mu, sd) # muestra con 100 observaciones, mu = 5 y sd = 1

# Calculamos la ecdf y luego la cdf
plot(ecdf(sample), main = "", ylab = "Probabilidad")
curve(pnorm(x, mean = mu, sd = sd), add = TRUE, col = "blue", lwd = 2)

# Calculamos la máxima distancia ("a mano")

sample_ordenada <- sort(sample)
Ui <- pnorm(sample_ordenada, mean = mu, sd = sd)
Dn_sup <- (1:n) / n - Ui
Dn_inf <- Ui - (1:n - 1) / n
i <- which.max(pmax(Dn_sup, Dn_inf))
lines(rep(sample_ordenada[i], 2),
      c(i / n, pnorm(sample_ordenada[i], mean = mu, sd = sd)),
      col = 4, lwd = 2)
rug(sample)
legend("topleft", lwd = 2, col = c(1:2, 4),
       legend = latex2exp::TeX(c("$F_n$", "$F_0$", "sup_x|F_n(x)-F_0(x)|")))
```



¿Como podriamos modificar la función `qnorm` para que rechacemos H_0?

Ejercicio 3

En este ejemplo haremos los calculos utilizando la función `ks.test` implementada en `R`. Se simula una muestra aleatoria para una distribución Normal con media 0 y desviación estandar de 1. La finalidad es determinar si rechazamos H_0 si se cree que $F_0 = F$

```
# Muestra aleatoria N(0,1)
set.seed(1000)            # Para que sea reproducible el ejercicio
n = 25
x = rnorm(n)

# Test de K-S para H_0: F = N(0, 1) (Rechazamos H0?)
(ks <- ks.test(x = x, y = "pnorm"))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.2156, p-value = 0.1686
## alternative hypothesis: two-sided
```

```
# Test de K-S para H_0: F = N(0.5, 1) (Rechazamos H0?)
ks.test(x = x, y = "pnorm", mean = 0.5)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.41288, p-value = 0.0002267
## alternative hypothesis: two-sided
```

```
# Test de K-S para H_0: F = Exp(2) (Rechazamos H0?)
ks.test(x = x, y = "pexp", rate = 1/2)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.69891, p-value = 9.308e-13
## alternative hypothesis: two-sided
```

Ejercicio 4

Considere que sólo se cuenta con 7 observaciones de una muestra aleatoria y se quiere probar que $F_0 = F$.

```
n = 7            # n observaciones
sim = 1000    # numero de simulaciones

# Creamos una variable que guarda las simulaciones
D = rep(0, sim)
for (j in 1 : sim) {

# Simulación d elas variables uniformes
x = runif(n, 0, 1)

# Ordenamos la muestra
x = sort(x)

# Calculo de la ecdf (funcion distribucion empirica)
Fn = ecdf(x)

# A la muestra ordenada le agregamos el 0 al inicio
# para cuando F_n(x(0))=0
y = c(0, x)

# Comenzamos con la busqueda del supremo
D1 = 0
D2 = 0
for (i in 2:(n+1)){
D1[i]= abs (Fn(y[i]) - y[i])
D2[i]= abs(Fn(y[i-1]) - y[i])
}

# Obtenemos maximo de maximos
D[j] = max(D1,D2)
if( j %% 1000 == 0) print(j)
}
```

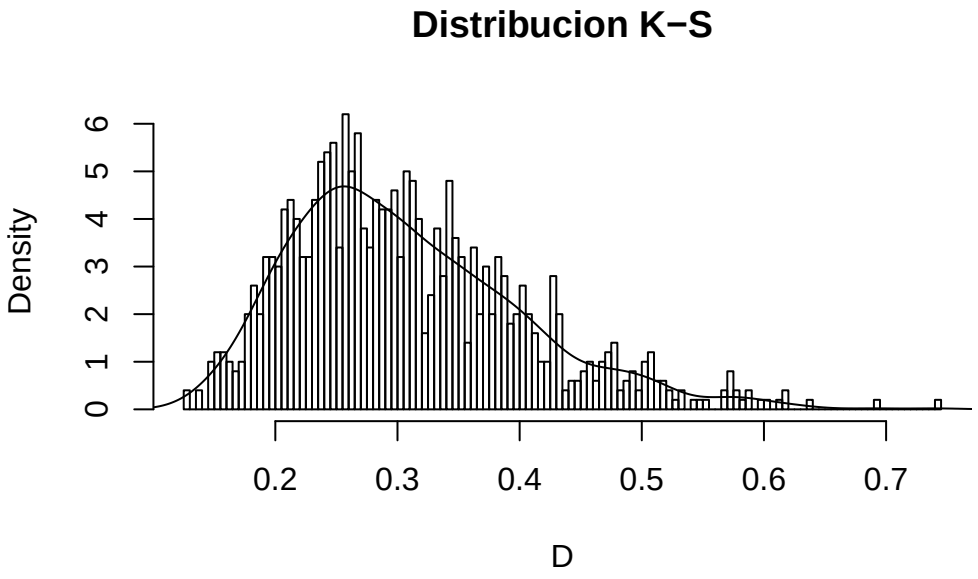
```
## [1] 1000
```

```
D.n = max(D1,D2)
D.n
```

```
## [1] 0.1997348
```

Ya creado el proceso de simulación y encontrado el estadístico D_n, gráficamos el hitograma que muestra la aproximación de la distribución K-S

```
hist(D, freq = FALSE, breaks = 100, main="Distribucion K-S")
lines(density(D))
```



Vemos los cuantiles estimados:

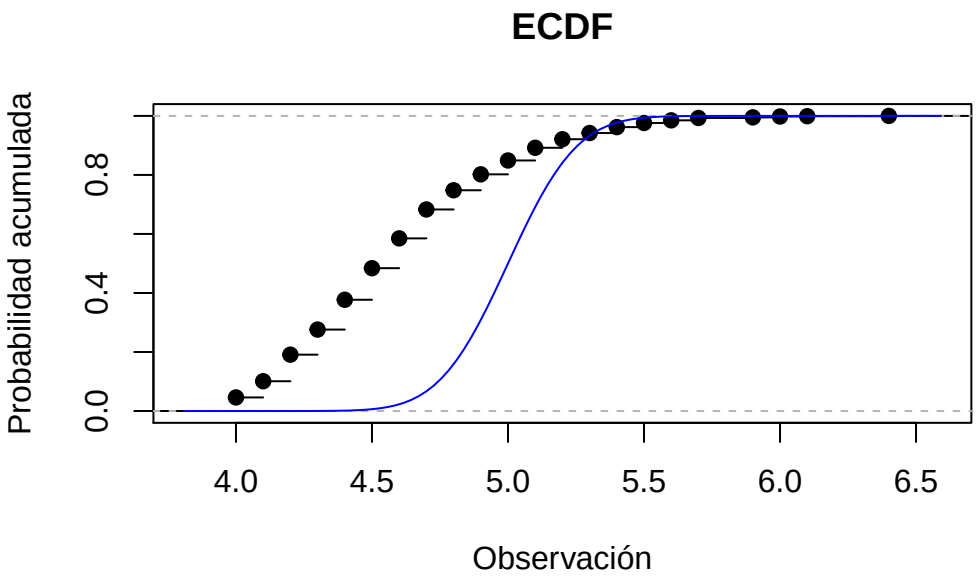
```
round(quantile(D, c(0.80,0.90,0.95,0.98,0.99)),3)
```

```
##   80%   90%   95%   98%   99%
## 0.382 0.431 0.489 0.547 0.582
```

Ejercicio 5

Se tienen los datos de temblores en la isla de Fiji medidos en magnitudes. Se quiere determinar si la muestra aleatoria proviene de una distribución Normal con media 5 y varianza 0.2. Los calculos para determinar el estadístico de prueba K-S son los siguientes:

```
# Directorio de trabajo
setwd("C:/Users/Usuario/Desktop/Lectures/non_parametrics/clases/clase4")

# Distribución de referencia
mu = 5
sigma = 0.2

### Le el conjunto de datos
x <- read.table("fijidata.txt", header=T)
x = x[, 5]

f = ecdf(x)
plot(f, xlab="Observación", main="ECDF", ylab="Probabilidad acumulada",
     col = "black")
curve(pnorm(x, mu, sigma),add=TRUE,col= "blue")
```



El contraste de hipótesis en este caso es el siguiente:

$$
\begin{aligned}
H_0 &: F_X(x) = \mathcal{N}(5,0{,}1) \\
H_0 &: F_X(x) \neq \mathcal{N}(5,0{,}1)
\end{aligned}
\tag{2}
$$

Primero calculamos el estadístico de prueba D_n:

```
# Directorio de trabajo
setwd("C:/Users/Usuario/Desktop/Lectures/non_parametrics/clases/clase4")

### Le el conjunto de datos
x = read.table("fijidata.txt", header=T)

# Seleccionamos la columna que nos interesa
x = x[, 5]

# Tamaños de la muestra
n = length(x)

# Ordenamos la muestra
x = sort(x)

# Calculamos la ecdf
Fn = ecdf(x)

# A la muestra ordenada le agregamos el 0 al inicio
# para cuando F_n(x(0))=0
y = c(0, x)

#Inicializamos busqueda de supremo
D1 = 0
D2 = 0
for (i in 2:(n+1)){
D1[i] = abs(Fn(y[i]) - pnorm(y[i], mu, sqrt(sigma)))
D2[i] = abs(Fn(y[i-1]) - pnorm(y[i], mu, sqrt(sigma)))
}
# Obtenemos el estadistico de prueba Dn
D.n = max(D1,D2)
D.n
```

```
## [1] 0.4318325
```

El estadístico de prueba D_n es igual a 0.4318325 el cual debemos comparar con el cuantil correspondiente cuando la distribución K-S es igual a 1000 (número de observaciones). Existen tablas en donde puede obtenerse este valor, pero como nosotros queremos evitar este paso (y sabemos el proceso de simulación asociado), haremos los calculos asociados a la obtención de este estadístico:

```
n = length(x)
sim = 1000    # numero de simulaciones

# Creamos una variable que guarda las simulaciones
D = rep(0, sim)
for (j in 1 : sim) {
```

```
# Simulación d elas variables uniformes
x = runif(n, 0, 1)

# Ordenamos la muestra
x = sort(x)

# Calculo de la ecdf (funcion distribucion empirica)
Fn = ecdf(x)

# A la muestra ordenada le agregamos el 0 al inicio
# para cuando F_n(x(0))=0
y = c(0, x)

# Comenzamos con la busqueda del supremo
D1 = 0
D2 = 0
for (i in 2:(n+1)){
D1[i]= abs (Fn(y[i]) - y[i])
D2[i]= abs(Fn(y[i-1]) - y[i])
}

# Obtenemos maximo de maximos
D[j] = max(D1,D2)
if( j %% 1000 == 0) print(j)
}
```

```
## [1] 1000
```

```
round(quantile(D, c(0.80,0.90,0.95,0.98,0.99)),3)
```

```
##   80%   90%   95%   98%   99%
## 0.034 0.039 0.043 0.046 0.050
```

El cuantil del 95 % de la distribución cuando es de $w_{0,95} = 0.043$, por lo tanto, como $D_n = 0{,}4318325 > 0{,}043$ rechazamos H_0.

Para comprobar podemos utilizar la función `ks.test` de `R`:

```r
set.seed(1000)
# Directorio de trabajo
setwd("C:/Users/Usuario/Desktop/Lectures/non_parametrics/clases/clase4")

### Le el conjunto de datos
x = read.table("fijidata.txt", header=T)

# Seleccionamos la columna que nos interesa
x = x[, 5]

# Tamaños de la muestra
n = length(x)

# Ordenamos la muestra
x = sort(x)
mu = 5
sigma = 0.2
ks.test(x, pnorm, mean = mu, sd = sigma, exact = TRUE)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.61619, p-value = 6.106e-15
## alternative hypothesis: two-sided
```

Test K-S de una cola

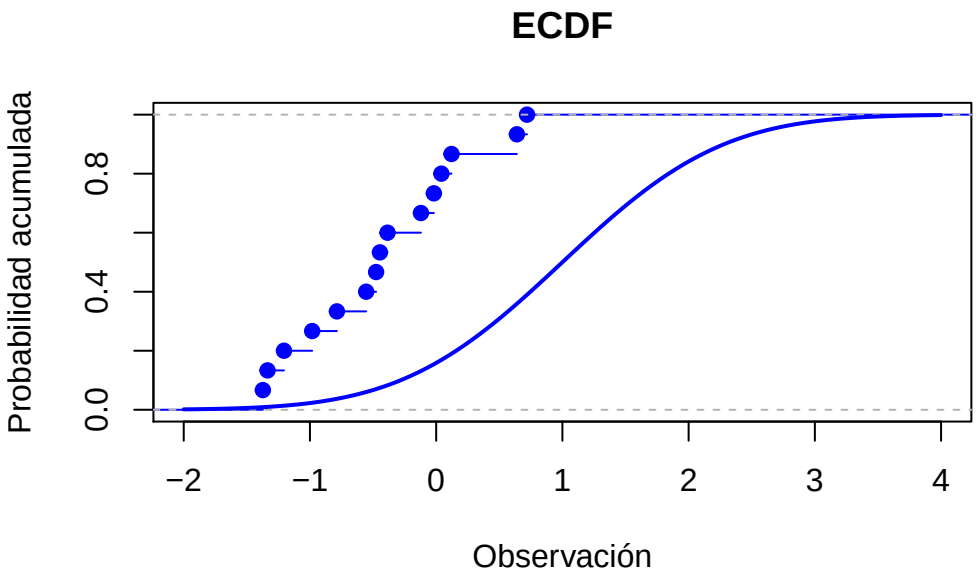
En este tipo de test no paramétrico también es posible definir pruebas de una sola cola, por lo que podemos plantear un contraste de hipótesis de la siguiente manera:

$$
\begin{aligned}
H_0 : F_X(x) \leq F_X^0(x) \\
H_0 : F_X(x) > F_X^0(x)
\end{aligned}
\tag{3}
$$

Para este caso el estadístico de prueba sólo considera la diferencia en un sentido sin tomar el valor absoluto de la ecuación de D_n. Para rechazar H_0 entonces debemos fijarnos que la verdadera distribución esta por encima de la distribución propuesta F_X^0.

Ejercicio 6

```
x = rnorm(15,0,1)
f = ecdf(x)
plot(f, xlab="Observación", xlim=c(-2,4), main="ECDF", ylab="Probabilidad acumulada",col=4)
curve(pnorm(x, 1, 1),add=TRUE,col="blue", lwd = 2)
```



Se puede apreciar en la figura que la ecdf siempre esta por encima de la distribución propuesta, por lo que en este caso particular se debe rechazar la hipótesis nula H_0. Si la ecdf esta sobrepasando por encima la distribución propuesta entonces podemos utilizar la siguiente ecuación:

$$
D_n^+ = \sup_{x \in \mathbb{R}} (F_n(x) - F_X^0(x)), \tag{4}
$$

por lo que debemos obtener la distribución de D_n^+ bajo H_0 mediante simulación con el fin de encontrar los cuantiles asociados cuando $D_n^+ > w_{1-\alpha}$. Para el caso inverso, en donde

$$\begin{aligned} H_0 &: F_X(x) \geq F_X^0(x) \\ H_0 &: F_X(x) < F_X^0(x) \end{aligned} \tag{5}$$

entonces el estadístico de prueba es:

$$D_n^- = \sup_{x \in \mathbb{R}} (F_X^0(x) - F_n(x)), \tag{6}$$

donde se rechaza la hipótesis nula H_0 si $D_n^- > w_{1-\alpha}$ con $w_{1-\alpha}$ es el cuantil $1-\alpha$ asociado a la distribución del estadístico D_n^- bajo H_0.