Ejercicios 5

Test de Kolmogorov-Smirnov (segunda parte)

Joaquin Cavieres G.

Introducción

Retomemos la idea detrás de aplicar pruebas de bondad de ajuste (test). Cuando la función de distribución bajo H_0 es continua nosotros podemos utilizar el método de bondad de ajuste X^2, sin embargo, nosotros debemos aproximar $F_0(x)$ mediante una categorización o agrupamiento de los datos observados a los cuales previamente les llamamos "clases" y denotamos por k. Para utilizar el test X^2 debemos tener un número grande de observaciones (un n "grande"), por lo que este test se encuentra limitado si consideramos una $F_0(x)$ continua pero con un tamaño de muestra pequeño. En cambio, el test K-S no necesita que los datos se categoricen o se agrupen en k clases y puede se aplicado sin problemas aún cuando el n observado sea pequeño.

El K-S se basa en la comparación de la función de distribución empírica acumulada (ECDF, por sus siglas en inglés) de la muestra ordenada y la función de distribución de referencia propuesta bajo H_0. Si la diferencia entre la ECDF ($F_n(x)$) y la función de distribución de referencia $F_0(x)$ es pequeña entonces generalmente no se rechaza H_0. Al contrario, si la diferencia entre $F_n(x)$ y $F_0(x)$ es grande, entonces rechazamos H_0.

Considere n puntos ordenados de datos observados $X_1, ..., X_n$. Dado lo anterior es que la ECDF, que denotamos como $F_n(x)$, puede calcularse como:

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

Para cualquier valor ordenado de x de la muestra aleatoria, $F_n(x)$ es la proporción del número de valores en la muestra que son iguales o menores a x. Como $F_0(x)$ esta completamente especificada (es la función de referencia), entonces nosotros podemos evaluar a $F_0(x)$ para algún valor de x y así comparar este x con el valor correspondiente a $F_n(x)$. Si H_0 es verdadera entonces la diferencia entre una y otra sea pequeña. De lo anterior podemos definir al estadístico K-S como:

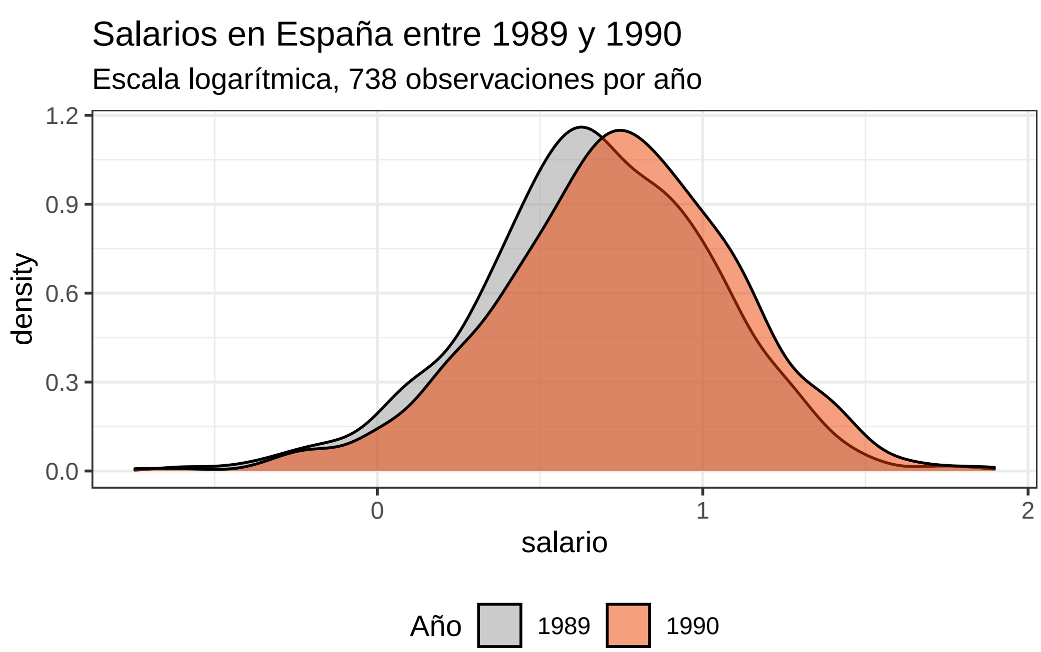
$$D_n = \sup_{x\in\mathbb{R}} \left|F_n(x) - F_0(x)\right|.$$

El estadístico D_n tiene una distribución que es independiente del modelo propuesto en la hipótesis nula (H_0), lo que nos permite evaluar la función de distribución de D_n sólo en función del tamaño de la muestra y en punto de $F_0(x)$.

Ejemplo 1

Utilizaremos el conjunto de datos `Snmesp` de la librería `plm` la cual contiene observaciones de los salarios (en escala `log`) en España durante los años 1983 y 1990 (un $n = 793$ para cada año). Nosotros queremos comprobar los salarios cambiaron entre 1989 y 1990.

```
library(tidyverse) # Cargamos la librería
#install.packages(plm)
library(plm)
data(Snmesp)      # Cargamos los datos

Snmesp <- Snmesp %>%
  dplyr::filter(year %in% c(1989, 1990)) %>%
  dplyr::mutate(year = as.factor(year)) %>%
  dplyr::select(year, salario = w)

Snmesp %>%
  ggplot(aes(x = salario, fill = year)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("gray60", "orangered2")) +
  labs(title = "Salarios en España entre 1989 y 1990",
       subtitle = "Escala logarítmica, 738 observaciones por año",
       fill = "Año") +
  theme_bw() +
  theme(legend.position = "bottom")
```



Ahora hacemos el calculo de la función de distribución empírica acumulada (ECDF) para nuestros datos observados (salarios). Esto lo podemos hacer a través de la función `ecdf` de `R` en donde recibe como argumento un vector de observaciones y devuelve como resultado una probabilidad acumulada.

```r
# Vemos el contenido de los datos
head(Snmesp, 6)
```

```
##   year   salario
## 1 1989  0.7769033
## 2 1990  1.0398120
## 3 1989 -0.3012976
## 4 1990 -0.2344583
## 5 1989  1.5129600
## 6 1990  1.6178640
```

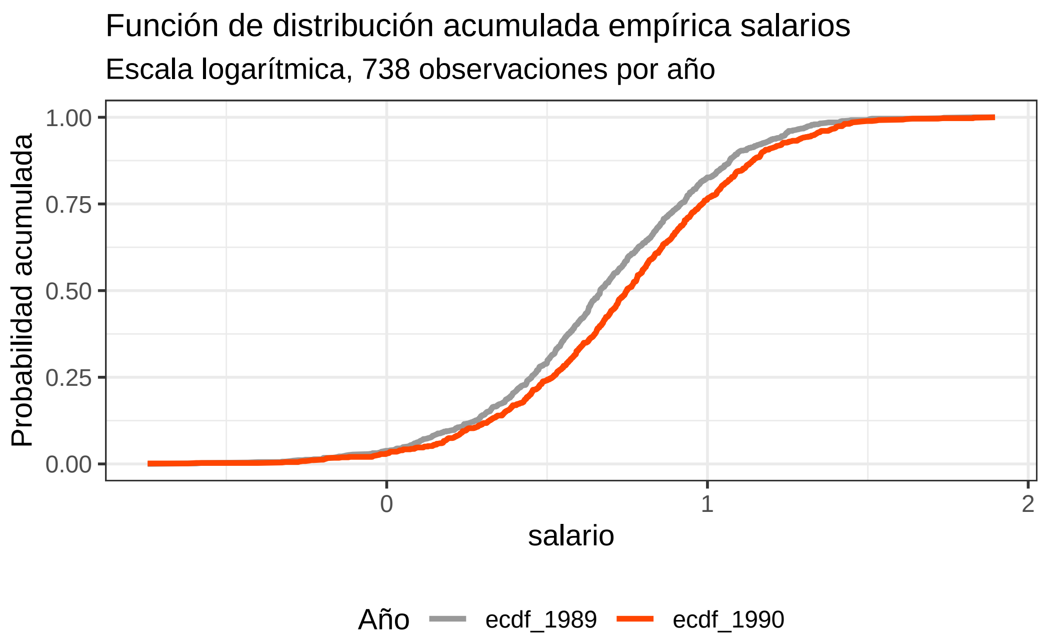
```r
summary(Snmesp)
```

```
##    year        salario
##  1989:738   Min.   :-0.7453
##  1990:738   1st Qu.: 0.4755
##             Median : 0.7102
##             Mean   : 0.7035
##             3rd Qu.: 0.9470
##             Max.   : 1.8965
```

```r
# Calculamos la ecdf para cada vector de años
ecdf_1989 = ecdf(Snmesp %>% filter(year == 1989) %>% pull(salario))
ecdf_1990 = ecdf(Snmesp %>% filter(year == 1990) %>% pull(salario))

# Se calcula la probabilidad acumulada de cada valor de salario observado con cada
# una de las funciones ecdf.
grid_salario <- unique(Snmesp %>% pull(salario))
prob_acumulada_ecdf_1989 = ecdf_1989(v = grid_salario)
prob_acumulada_ecdf_1990 = ecdf_1990(v = grid_salario)
```

```r
# Se ajustan las funciones ecdf con cada muestra.
# Se unen los valores calculados en un dataframe.
data_ecdf = data.frame(salario = grid_salario,
            ecdf_1989 = prob_acumulada_ecdf_1989,
            ecdf_1990 = prob_acumulada_ecdf_1990) %>%
            pivot_longer(
            cols = c(ecdf_1989, ecdf_1990),
            names_to = "year",
            values_to = "ecdf")

plot_ecdf = ggplot(data = data_ecdf,
               aes(x = salario, y = ecdf, color = year)) +
               geom_line(size = 1) +
               scale_color_manual(values = c("gray60", "orangered1")) +
```

```
                 labs(title = "Función de distribución acumulada empírica salarios",
                 subtitle = "Escala logarítmica, 738 observaciones por año",
                 color = "Año",
                 y = "Probabilidad acumulada") +
                 theme_bw() +
                 theme(legend.position = "bottom",
                 plot.title = element_text(size = 12))

plot_ecdf
```



Ahora realizamos el test K-S. Este calculo lo haremos "a mano":

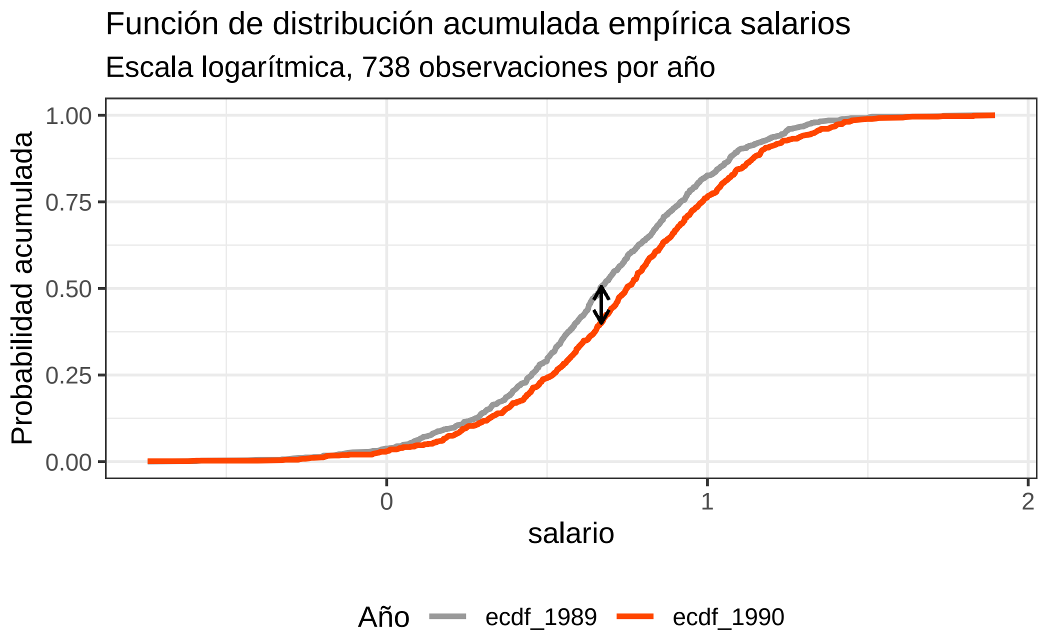
```
# Se calcula la diferencia absoluta entre las probabilidades acumuladas de cada
# función.
abs_dif <-  abs(prob_acumulada_ecdf_1989 - prob_acumulada_ecdf_1990)

# La distancia Kolmogorov-Smirnov es el máximo de las distancias absolutas.
D_n <- max(abs_dif)
paste("Distancia Kolmogorov-Smirnov:", D_n)
```

```
## [1] "Distancia Kolmogorov-Smirnov: 0.105691056910569"
```

y agregamos la distancia calculada al gráfico de las ECDF's calculadas previamente:

```
indice_ks <- which.max(abs_dif)
```

```r
plot_ecdf + geom_segment(aes(x = grid_salario[indice_ks],
                xend = grid_salario[indice_ks],
                y = prob_acumulada_ecdf_1989[indice_ks],
                yend = prob_acumulada_ecdf_1990[indice_ks]),
               arrow = arrow(ends = "both", length = unit(0.2,"cm")),
               color = "black")
```



Para validar nuestro calculo compararemos nuestro D_n estimado con el estimado mediante la función `ks.test()` disponible en `R`:

```r
test_ks = ks.test(
        x = Snmesp %>% filter(year == 1989) %>% pull(salario),
        y = Snmesp %>% filter(year == 1990) %>% pull(salario))
test_ks$statistic
```

```
##          D
## 0.1056911
```

Para este ejemplo en particular, ¿rechazamos H_0?

Ya habiendo calculado D_n debemos determinar si este valor de esta distancia entre $F_n(x)$ y $F_0(x)$ es suficientemente grande considerando la muestra aleatoria (datos observados) para así determinar si las dos distribuciones son distintas (basados en el p-value). Esto podemos determinarlo facílmente mediante la función `ks.test()`:

```
test = ks.test(
        x = Snmesp %>% filter(year == 1989) %>% pull(salario),
        y = Snmesp %>% filter(year == 1990) %>% pull(salario))
test
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  Snmesp %>% filter(year == 1989) %>% pull(salario) and Snmesp %>% filter(year == 1990)
## D = 0.10569, p-value = 0.0005257
## alternative hypothesis: two-sided
```

Existen evidencias empíricas para considerar que la distribución de salarios para el año 1989 y 1900 no son las mismas.

Bandas de confianza para $F_0(x)$

Al conocer la función de distribución del estadístico $D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|$ nosotros podemos encontrar las bandas de confianza de la verdadera función de distribución $F_0(x)$. Denotamos a $w_{1-\alpha}$ como el cuantil $1-\alpha$ de la distribución del estadístico D_n, entonces:

$$
\begin{aligned}
\mathbb{P}(D_n \leq w_{1-\alpha}) &= 1-\alpha \\
\mathbb{P}(\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| \leq w_{1-\alpha}) &= 1-\alpha \\
\mathbb{P}(|F_n(x) - F_0(x)| \leq w_{1-\alpha}\ \ \forall x \in \mathbb{R}) &= 1-\alpha \\
\mathbb{P}(-w_{1-\alpha} \leq F_n(x) - F_0(x) \leq w_{1-\alpha}\ \ \forall x \in \mathbb{R}) &= 1-\alpha \\
\mathbb{P}(F_n(x) - w_{1-\alpha} \leq F_0(x) \leq F_n(x) + w_{1-\alpha}\ \ \forall x \in \mathbb{R}) &= 1-\alpha
\end{aligned}
$$

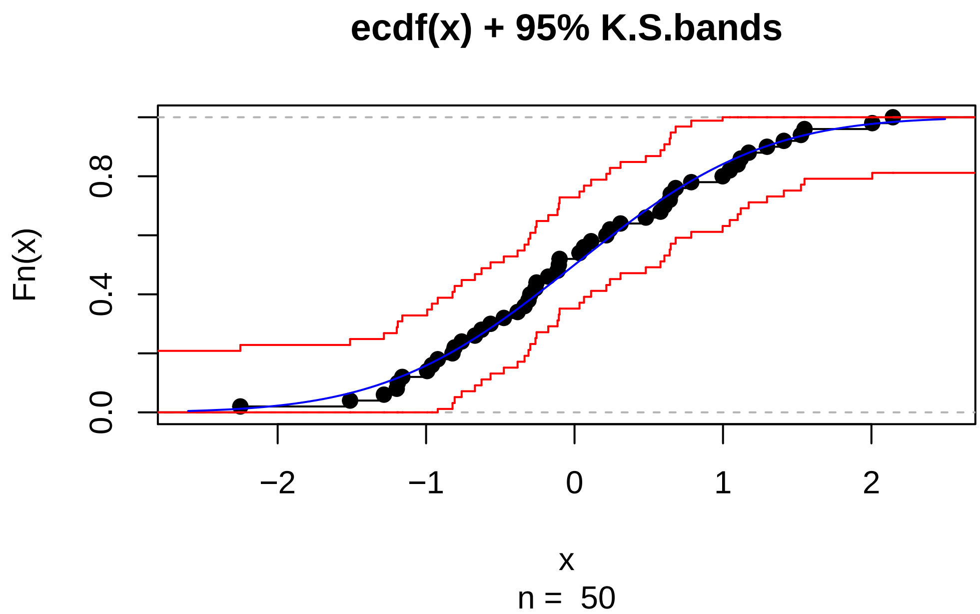
Así $F_n(x) - w_{1-\alpha}$ y $F_n(x) + w_{1-\alpha}$ forman la banda de confianza para $F_0(x)$.

Este calculo se puede realizar directamente en `R` con la librería `NSM3` y la función `ecdf.ks.CI`:

```r
#install.packages("NSM3")
library(NSM3)
# Simulamos 50 observaciones desde una dist Normal estandar
x = rnorm(50,0,1)
# Consturimos las bandas de confianza
ecdf.ks.CI(x)
```

```
## $lower
##  [1] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
## [10] 0.01159 0.03159 0.05159 0.07159 0.09159 0.11159 0.13159 0.15159 0.17159
## [19] 0.19159 0.21159 0.23159 0.25159 0.27159 0.29159 0.31159 0.33159 0.35159
## [28] 0.37159 0.39159 0.41159 0.43159 0.45159 0.47159 0.49159 0.51159 0.53159
## [37] 0.55159 0.57159 0.59159 0.61159 0.63159 0.65159 0.67159 0.69159 0.71159
## [46] 0.73159 0.75159 0.77159 0.79159 0.81159
##
## $upper
##  [1] 0.20841 0.22841 0.24841 0.26841 0.28841 0.30841 0.32841 0.34841 0.36841
## [10] 0.38841 0.40841 0.42841 0.44841 0.46841 0.48841 0.50841 0.52841 0.54841
## [19] 0.56841 0.58841 0.60841 0.62841 0.64841 0.66841 0.68841 0.70841 0.72841
## [28] 0.74841 0.76841 0.78841 0.80841 0.82841 0.84841 0.86841 0.88841 0.90841
## [37] 0.92841 0.94841 0.96841 0.98841 1.00000 1.00000 1.00000 1.00000 1.00000
## [46] 1.00000 1.00000 1.00000 1.00000 1.00000
```

```r
# Agregamos la curva teorica de donde vino la muestra
curve(pnorm(x,0,1), add = TRUE, col = "blue")
```


ecdf(x) + 95% K.S.bands
Fn(x)
0.0
0.4
0.8
−2
−1
0
1
2
x
n = 50