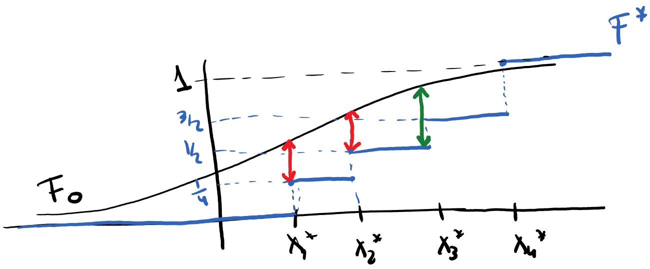
Práctico 12

martes, 21 de junio de 2022

# Prueba de Kolmogorov - Smirnov

$X_1, \dots, X_n$ MAS de $X$, no conocemos la distribución de $X$

$$\begin{cases} H_0: X_1, \dots, X_n \text{ tienen distribución } F_0 \\ H_1: X_1, \dots, X_n \text{ no tienen distribución } F_0 \end{cases}$$




ordenamos los datos: $X_1^* \leq X_2^* \leq X_3^* < \dots < X_n^*$

$$F^*(X_i^*) = \frac{i \text{ casos favorables}}{n \text{ casos posibles}} = \frac{i}{n}$$

$$RC = \{ \sup_{x \in \mathbb{R}} |F_0(x) - F^*(x)| \geq \underbrace{c(\alpha, n)}_{\downarrow} \}$$

tabla de Kolmogorov - Smirnov

resultado:

$$\sup_{x \in \mathbb{R}} |F_0(x) - F^*(x)| = \max \{ \underline{\left| i/n - F_0(X_i^*) \right|}, \underline{\left| \frac{i-1}{n} - F_0(X_i^*) \right|} \}$$

Procedimiento:

| $i$ | $X_i^*$ | $\left\lvert \frac{i}{n} - F_0(X_i^*) \right\rvert$ | $\left\lvert \frac{i-1}{n} - F_0(X_i^*) \right\rvert$ |
|---|---|---|---|
| 1 | | | |
| 2 | | | |
| 3 | | | |
| $\vdots$ | | | |

buscamos el valor máximo y lo comparamos con $c(\alpha,n)$

**Ejercicio 1**

Para la siguiente muestra

| | | | | |
|---|---|---|---|---|
| 0.073 | 0.021 | 0.162 | 0.094 | 0.303 |
| 0.018 | 0.08 | 0.061 | 0.19 | 0.079 |

Realizar la prueba de Kolmogorov y Smirnov para ver si es razonable afirmar que los datos tienen distribución exponencial de parámetro $\lambda = 2$.

$$\begin{cases} H_0: \text{los datos tienen distribución } \exp(2) \quad F_0(x) = 1 - e^{-2x} \\ H_1: \text{los datos no tienen distribución } \exp(2) \end{cases}$$

$$RC = \left\{ \sup_{x \in \mathbb{R}} \left\lvert F_0(x) - F^*(x) \right\rvert \geq c(\alpha, n) \right\}$$

$$\left.\begin{array}{l} \alpha = 0{,}05 \\ n = 10 \end{array}\right\} \Rightarrow c(\alpha,n) = 0{,}410$$

**Tabla 8 Contraste de Kolmogorov-Smirnov**

Valores críticos de $D = |F_n(x) - F(x)|$ donde $F_n(x)$ es la distribución muestral de tamaño $n$ y $F(x)$ la distribución teórica.

| Tamaño muestral $n$ | Nivel de significación | | | | |
|---|---|---|---|---|---|
| | 0,20 | 0,15 | 0,10 | 0,05 | 0,01 |
| 1 | 0,900 | 0,925 | 0,950 | 0,975 | 0,995 |
| 2 | 0,684 | 0,726 | 0,776 | 0,842 | 0,929 |
| 3 | 0,565 | 0,597 | 0,642 | 0,708 | 0,828 |
| 4 | 0,494 | 0,525 | 0,564 | 0,624 | 0,733 |
| 5 | 0,446 | 0,474 | 0,510 | 0,565 | 0,669 |
| 6 | 0,410 | 0,436 | 0,470 | 0,521 | 0,618 |
| 7 | 0,381 | 0,405 | 0,438 | 0,486 | 0,577 |
| 8 | 0,358 | 0,381 | 0,411 | 0,457 | 0,543 |
| 9 | 0,339 | 0,360 | 0,388 | 0,432 | 0,514 |
| 10 | 0,322 | 0,342 | 0,368 | $\boxed{0,410}$ | 0,490 |
| 11 | 0,307 | 0,326 | 0,352 | 0,391 | 0,468 |
| 12 | 0,295 | 0,313 | 0,338 | 0,375 | 0,450 |

| $i$ | $X_i^*$ | $\left\| \frac{i}{10} - (1-e^{-2X_i^*}) \right\|$ | $\left\| \frac{i-1}{10} - (1-e^{-2X_i^*}) \right\|$ |
|---|---|---|---|
| 1 | 0,018 | 0,065 | |
| 2 | 0,021 | 0,159 | |
| 3 | 0,061 | 0,185 | |
| 4 | 0,073 | 0,264 | |
| 5 | 0,079 | 0,354 | |
| 6 | 0,08 | 0,452 | |
| 7 | 0,094 | 0,529 | |
| 8 | 0,162 | 0,524 | |
| 9 | 0,19 | 0,584 | |
| 10 | 0,303 | 0,546 | |

$$\sup_{x \in \mathbb{R}} |F_0(x) - F^*(x)| \geq 0,452 > 0,410$$

entonces la muestra está en la región crítica

$\Rightarrow$ rechazamos $H_0$

**Ejercicio 3**
Para la siguiente muestra

| | | | | |
|---|---|---|---|---|
| 6.586 | 1.706 | 1.437 | 6.62 | 10.849 |
| 9.709 | 10.352 | 6.48 | 12.318 | 7.198 |

1. Realizar la prueba de Kolmogorov y Smirnov para ver si es razonable afirmar que los datos tienen distribución normal con media $\mu = 3$ y desvío $\sigma = 2$.
2. Realizar la prueba de Kolmogorov-Smirnov-Lilliefors para ver si es razonable afirmar que los datos tienen distribución normal.

1) $H_0$: los datos tienen distribución $\mathcal{N}(3, 2^2)$    $F_0$ es la función de distribución de $\mathcal{N}(3, 2^2)$

$H_1$: los datos no tienen distribución $\mathcal{N}(3, 2^2)$

$$RC = \left\{ \sup_{x \in \mathbb{R}} |F_0(x) - F^*(x)| \geq c(\alpha, n) \right\}$$

$\alpha = 0{,}05$ 
$n = 10$ $\Big\}$ en la tabla: $c(\alpha, n) = 0{,}410$

| $i$ | $X_i^*$ | $\left\vert \frac{i}{n} - \phi\left(\frac{X_i^* - 3}{2}\right) \right\vert$ | $\left\vert \frac{i-1}{n} - \phi\left(\frac{X_i^* - 3}{2}\right) \right\vert$ |
|---|---|---|---|
| 1 | 1,437 | 0,1173 | |
| 2 | 1,706 | 0,0588 | |
| 3 | 6,48 | $\boxed{0{,}6591}$ | |
| 4 | 6,586 | | |
| 5 | 6,62 | | |
| 6 | 7,198 | | |
| 7 | 9,709 | | |
| 8 | 10,352 | | |
| 9 | 10,842 | | |
| 10 | 12,318 | | |

$$\frac{i}{n} - F_0(x_i^*) = \frac{i}{n} - P(X \leq x_i^*) \quad \text{donde } X \sim \mathcal{N}(3, 2^2)$$

$$= \frac{i}{n} - P\left(\frac{X-3}{2} \leq \frac{x_i^* - 3}{2}\right)$$

$$= \frac{i}{n} - \phi\left(\frac{x_i^* - 3}{2}\right)$$

$$\sup_{x \in \mathbb{R}} |F_0(x) - F^*(x)| \geq 0{,}6591 > 0{,}410$$

entonces la muestra está en la región crítica

$\Rightarrow$ rechazamos $H_0$.

② rechazamos que los datos tengan distribución $\mathcal{N}(3, 2^2)$
pero podrían tener distribución normal con otros parámetros

consideramos $\begin{cases} \mu = \overline{X}_n \\ \sigma = s_n \end{cases}$

$\begin{cases} H_0: \text{los datos tienen distribución } \mathcal{N}(\overline{X}_n, s_n^2) \\ H_1: \text{los datos no tienen distribución } \mathcal{N}(\overline{X}_n, s_n^2) \end{cases}$

$F_0$ es la función de distribución de $\mathcal{N}(\overline{X}_n, s_n^2)$

$$RC = \left\{ \sup_{x \in \mathbb{R}} |F_0(x) - F^*(x)| \geq \underbrace{d(\alpha, n)}_{\downarrow} \right\}$$

tabla Kolmogorov-Smirnov-Lilliefors

$\left.\begin{array}{l} \alpha = 0{,}05 \\ n = 10 \end{array}\right\} \Rightarrow$ en la tabla: $d(\alpha, n) = 0{,}258$

**Tabla de Lilliefors**, para ajuste a la normal con parámetros estimados (no confundir con la del ajuste exponencial)

Tablas de $D_n = |F_n(x) - F(x)|$ para contrastar la hipótesis de normalidad cuando la media y la varianza poblacionales son estimadas por sus valores muestrales.

| Tamaño muestral $n$ | Nivel de significación | | | | |
|---|---|---|---|---|---|
| | 0,20 | 0,15 | 0,10 | 0,05 | 0,01 |
| 4 | 0,300 | 0,319 | 0,352 | 0,381 | 0,417 |
| 5 | 0,285 | 0,299 | 0,315 | 0,337 | 0,405 |
| 6 | 0,265 | 0,277 | 0,294 | 0,319 | 0,364 |
| 7 | 0,247 | 0,258 | 0,276 | 0,300 | 0,348 |
| 8 | 0,233 | 0,244 | 0,261 | 0,285 | 0,331 |
| 9 | 0,223 | 0,233 | 0,249 | 0,271 | 0,311 |
| 10 | 0,215 | 0,224 | 0,239 | $\boxed{0{,}258}$ | 0,294 |
| 11 | 0,206 | 0,217 | 0,230 | 0,249 | 0,284 |
| 12 | 0,199 | 0,212 | 0,223 | 0,242 | 0,275 |
| 13 | 0,190 | 0,202 | 0,214 | 0,234 | 0,268 |
| 14 | 0,183 | 0,194 | 0,207 | 0,227 | 0,261 |

| $i$ | $X_i^*$ | $\left\| \frac{i}{n} - \phi\left(\frac{X_i^* - \bar{X}_n}{s_n}\right)\right\|$ | $\left\| \frac{i-1}{n} - \phi\left(\frac{X_i^* - \bar{X}_n}{s_n}\right)\right\|$ |
|---|---|---|---|
| 1 | 1,437 | | |
| 2 | 1,706 | | $\boxed{0{,}2085}$ ← maximo de los valores de las dos columnas |
| 3 | 6,48 | | |
| 4 | 6,586 | | |
| 5 | 6,62 | | |
| 6 | 7,198 | | |
| 7 | 9,709 | | |
| 8 | 10,352 | | |
| 9 | 10,842 | | |
| 10 | 12,318 | | |

$$\frac{i}{n} - F_0(X_i^*) = \frac{i}{n} - P(X \leq X_i^*) \quad \text{donde } X \sim \mathcal{N}(\bar{X}_n, s_n^2)$$

$$= \frac{i}{n} - P\left(\frac{X - \bar{X}_n}{s_n} \leq \frac{X_i^* - \bar{X}_n}{s_n}\right)$$

$$= \frac{i}{n} - \phi\left(\frac{X_i^* - \bar{X}_n}{s_n}\right)$$

$$\sup_{x \in \mathbb{R}} |F_0(x) - F^*(x)| = 0{,}2085 < 0{,}258$$

$\Rightarrow$ la muestra no está en la región crítica

$\Rightarrow$ no rechazamos $H_0$.

**Ejercicio 6 (5 puntos)**

Una moneda tiene probabilidad de cara igual a $\theta$. Se desea hacer el siguiente test de hipótesis sobre el valor de $\theta$:

$$\begin{cases} H_0 : \theta = 0.4 \\ H_A : \theta = 0.7 \end{cases}$$

Para esto se la lanza 12 veces, y se usa como estadístico la cantidad de caras $X$.

A continuación se muestra la f.p.p. de una $\text{Bin}(12, \theta)$ para los dos valores de $\theta$:

$X \sim \text{Bin}(12, \theta)$

$RC = \{X \geq c\}$

$\alpha = P_{H_0}(X \geq c)$

Función de probabilidad puntual $p(x;\theta)$

| $\theta$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.4 | 0.002 | 0.017 | 0.064 | 0.142 | 0.213 | 0.227 | 0.177 | 0.101 | 0.042 | 0.012 | 0.003 | 0.000 | 0.000 |
| 0.7 | 0.000 | 0.000 | 0.000 | 0.002 | 0.008 | 0.029 | 0.079 | 0.158 | 0.231 | 0.240 | 0.168 | 0.071 | 0.014 |

Se toma como región de rechazo $\{X \geq c\}$, en donde $c$ es el menor entero entre 0 y 12 que cumple $P(X \geq c | H_0) \leq 0.1$.

Entonces, la potencia del test es:

**(A)** 0.057 **(C)** 0.010 **(E)** 0.276

**(B)** 0.900 **(D)** 0.724 **(F)** 0.158

buscamos c tal que c es el menor entero entre 0 y 12

$P_{H_0}(X \geq c) \leq 0,1$