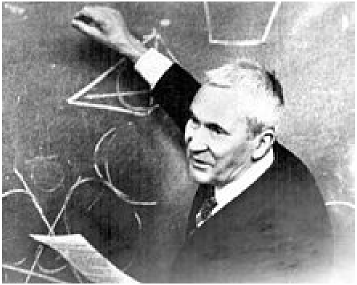


Figura: Andrei Kolmogorov (1903-1987, Rusia)

Clase 20 de Bioestadística

Test de Kolmogorov-Smirnov

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Test de distribución

Test de Kolmogorov

Test de dos muestras

Test de distribución

- Algunos procedimientos estadísticos se basan en conocer la distribución de los datos
- ¿Qué podemos hacer si disponemos de una muestra aleatoria simple de los datos?
- Suponemos aquí que tenemos una hipótesis para testear.
- Es decir, suponemos que los datos provienen de una cierta distribución F_0, que podría ser una uniforme, normal, exponencial, u otra.
- Escribimos entonces

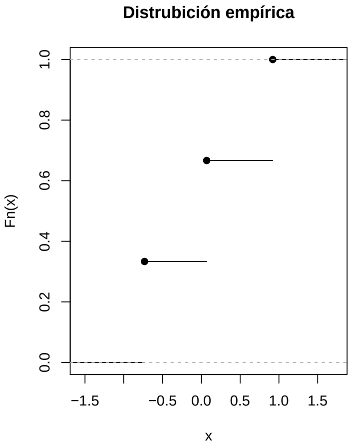
$$H_0 : F = F_0$$

- ¿Cómo construímos un estadístico para este test?

Distribución empírica

- Kolmogorov propuso utilizar la distribución empírica para compararla con la distribución F_0
- La distribución empírica es la escalera que sube $1/n$ en el lugar de cada dato
- Si la muestra es $X_1, \ldots, X_n$, la fórmula es

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k \leq x\}}$$



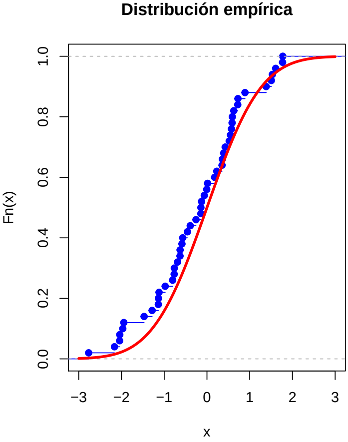
- Tenemos la distribución empírica con $n = 3$
- Son 3 escalones de altura 1/3

- Vimos que la distribución empírica converge a la distribución teórica, es decir

$$F_n(x) \to F_0(x)$$

(en el supuesto de que F_0 es cierta)

- La idea es medir la discrepancia o diferencia entre F_n y F_0.



- En azul la distribución empírica de 50 datos simulados
- La distribución en rojo es la teórica
- ¿Cómo medimos la distancia entre las dos funciones?

- La propuesta para medir la distancia entre ambas curvas es tomar la máxima diferencia en valor absoluto entre los valores que toma la función

- Es decir, calcular

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

- Observemos que medimos el punto en donde las funciones se diferencian mas.

- Se puede definir la distancia para dos funciones de distribución cualesquiera:

$$D(F, G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|$$

- Observamos que si $D(F, G) = 0$, entonces $F(x) = G(x)$ para todo x, es decir, son iguales

Test de Kolmogorov

- Se trata de calcular la distancia entre F_n y F_0 y decidir si es lo suficientemente grande como para rechazar la hipótesis nula.
- El criterio de rechazo es si cae en una zona de probabilidad baja ($\alpha = 0{,}05$)
- Kolmogorov encontró la forma de calcular probabilidades para esa distancia
- Introdujo la variable aleatoria $\mathcal{K}$ que cumple

$$P(\mathcal{K} \geq x) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2}$$

- Estas probabilidades están programadas en el R

Kolmogorov demostró un Teorema Central del Límite para esta situación:

Teorema

Supongamos que la distribución F tiene distribución continua. Entonces

$$K_n := \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{d} \mathcal{K}.$$

Esto quiere decir que

$$\mathbf{P}(K_n \leq x) \to \mathbf{P}(K \leq x).$$

Notablemente, la distribución límite no depende de F (igual que en el TCL)

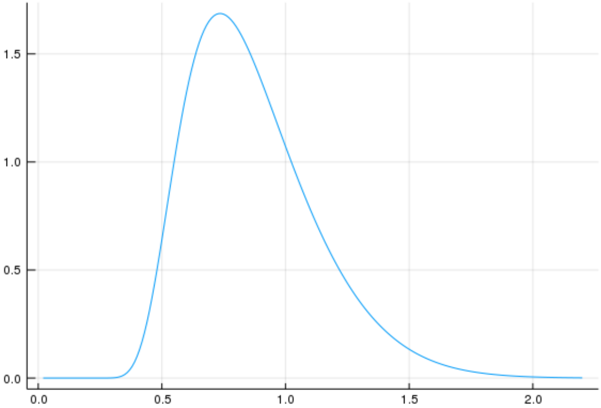


Figura: Densidad de la variable $\mathcal{K}$ de Kolmogorov

Test de hipótesis

Para testear la distribución de una muestra (por ejemplo, un generado aleatorio) hacemos un test de hipótesis mediante los siguientes pasos:

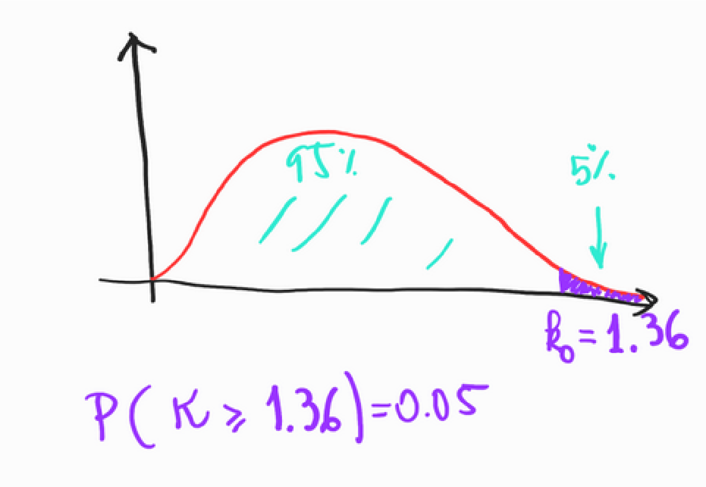
- Simulamos la muestra $(X_1, \dots, X_n)$ de una distribución F y calculamos K_n.
- Construimos un intervalo de confianza para K de nivel $1 - \alpha$, de la forma

$$\mathbf{P}(\mathcal{K} \geq k_0) = \alpha.$$

- Si nuestro valor calculado K_n es mayor que k_0, rechazamos la hipótesis nula de que las variables tienen distribución F_0

Observese que se testea F_0, pero se asume que las variables son independientes y tienen la misma distribución.

La distribución de Kolmogorov:



Ejemplo: distribución uniforme

- Supongamos que corremos el comando `runif(100)`
- Eso nos genera valores de una m.a.s. uniforme en $[0, 1]$ de tamaño 100
- Calculo
$$D_{100} = 0{,}081237$$
- Tengo $\sqrt{100}D = 0{,}81$
- El cuantil 0.95 de $\mathcal{K}$ es 1.36
- ¿Cuál es la conclusión?

- No rechazamos la hipótesis nula
- Este procedimiento se puede hacer en R con el comando `ks.test`

```
muestra <- runif(100)
```

```
> ks.test(muestra,"punif")

        One-sample Kolmogorov-Smirnov test

data:  muestra
D = 0.081237, p-value = 0.5242
alternative hypothesis: two-sided
```

Test de dos muestras

La distribución de Kolmogorov se usa también en la siguiente situación:

- Tenemos dos muestras aleatorias simples, independientes entre ellas.
- La primera muestra tiene n elementos:

$$X_1, X_2, \ldots, X_n$$

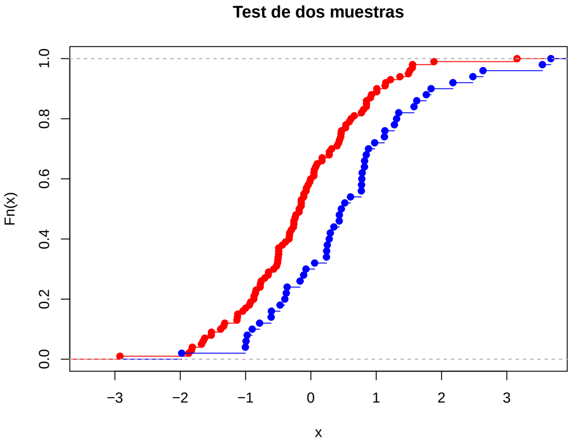
- La segunda muestra tiene m elementos

$$Y_1, \ldots, Y_m$$

- La hipótesis nula es que ambas muestras provienen de la misma distribución:

$$H_0 \colon F_X = F_Y$$

La idea es comparar ambas distribuciones empíricas (ambas escaleras)



El Teorema Central del Límite para esta situación es:

Teorema

Supongamos que la distribuciónes F_X y F_Y son continuas. Entonces

$$K_{m,n} := \sqrt{\frac{mn}{m+n}} \sup_{x \in \mathbb{R}} |F_n(x) - F_m(x)| \xrightarrow{d} \mathcal{K}.$$

Esto quiere decir que

$$\mathbf{P}(K_{m,n} \leq x) \to \mathbf{P}(K \leq x).$$

Notablemente, la distribución límite no depende de F (igual que en el TCL)

Ejemplo: un ensayo clínico (clinical trial)

- Se aplica un medicamento a un grupo de 50 personas y un placebo a otro grupo de 50 personas de las mismas características.
- Una variable de interés clínico sobre la que supuestamente actúa la droga se mide en ambos grupos
- Tenemos entonces dos muestras con $m = n = 50$
- Si el medicamento no tiene efecto, va a resultar $F_X = F_Y$ es decir, no observamos diferencias entre los grupos

- Supongamos que obtenemos

$$D = 0{,}32$$

- Calculamos

$$\sqrt{\frac{mn}{m+n}} = \sqrt{\frac{n^2}{2n}} = \sqrt{\frac{n}{2}} = \sqrt{25}$$

- Entonces

$$K_{m,n} = \sqrt{25} \times 0{,}32 = 1{,}6$$

- Como $k_0 = 1{,}36$ a nivel $\alpha = 0{,}05$ rechazo la hipótesis nula
- La diferencia estadística es suficiente para afirmar que el medicamento es efectivo.