Karl Pearson (1857-1936, UK)

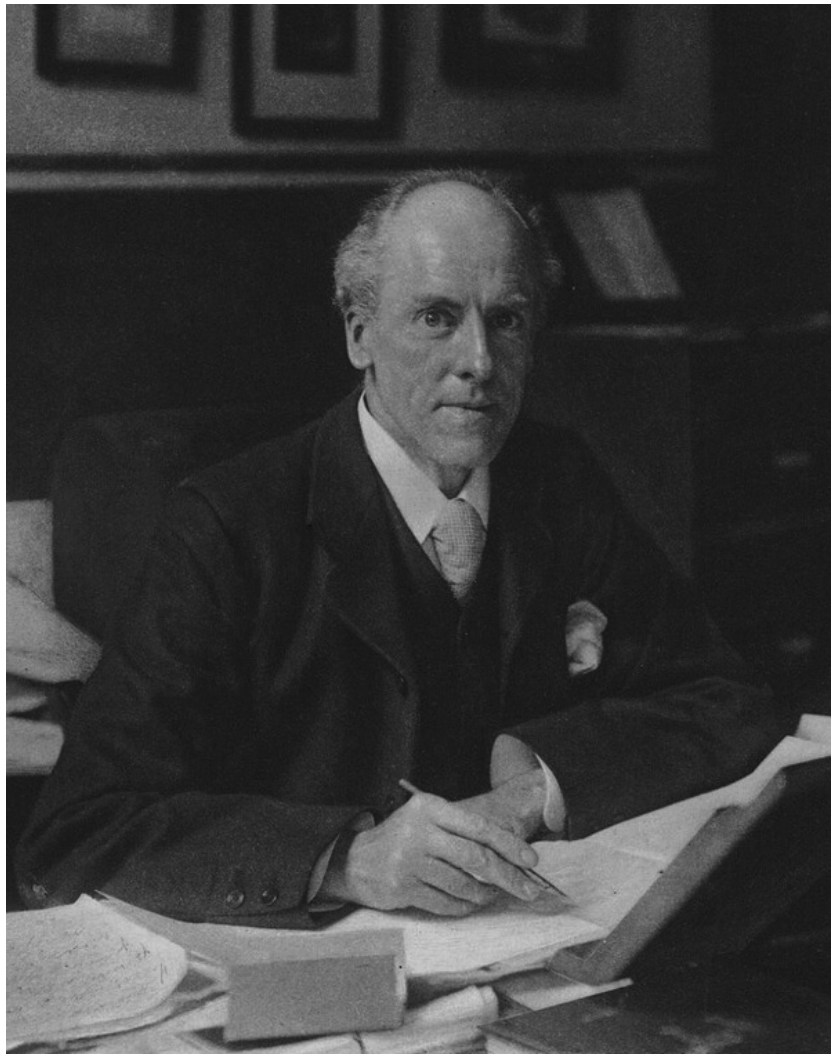


Figura: Karl Pearson en 1910. Es considerado uno de los fundadores de la bioestadística

Clase 21 de Bioestadística

Test de Lilliefors y Chi-cuadrado de Pearson

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

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

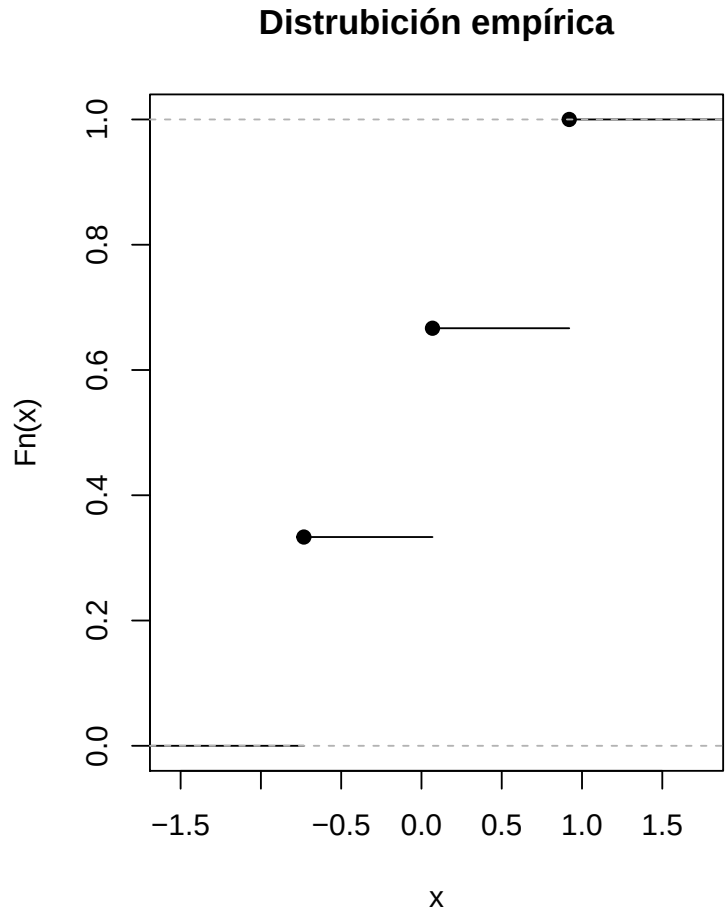
$$H_0 : F = F_0$$

- ¿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

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



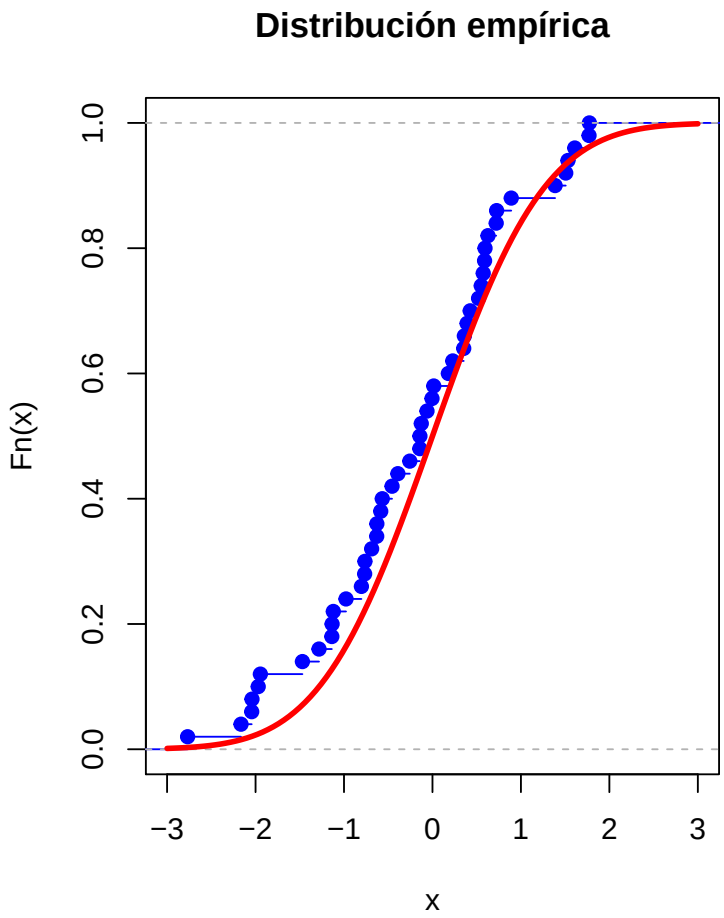
- Tenemos la distribución empírica con $n = 3$
- Son 3 escalones de altura 1/3

- 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

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

- 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)

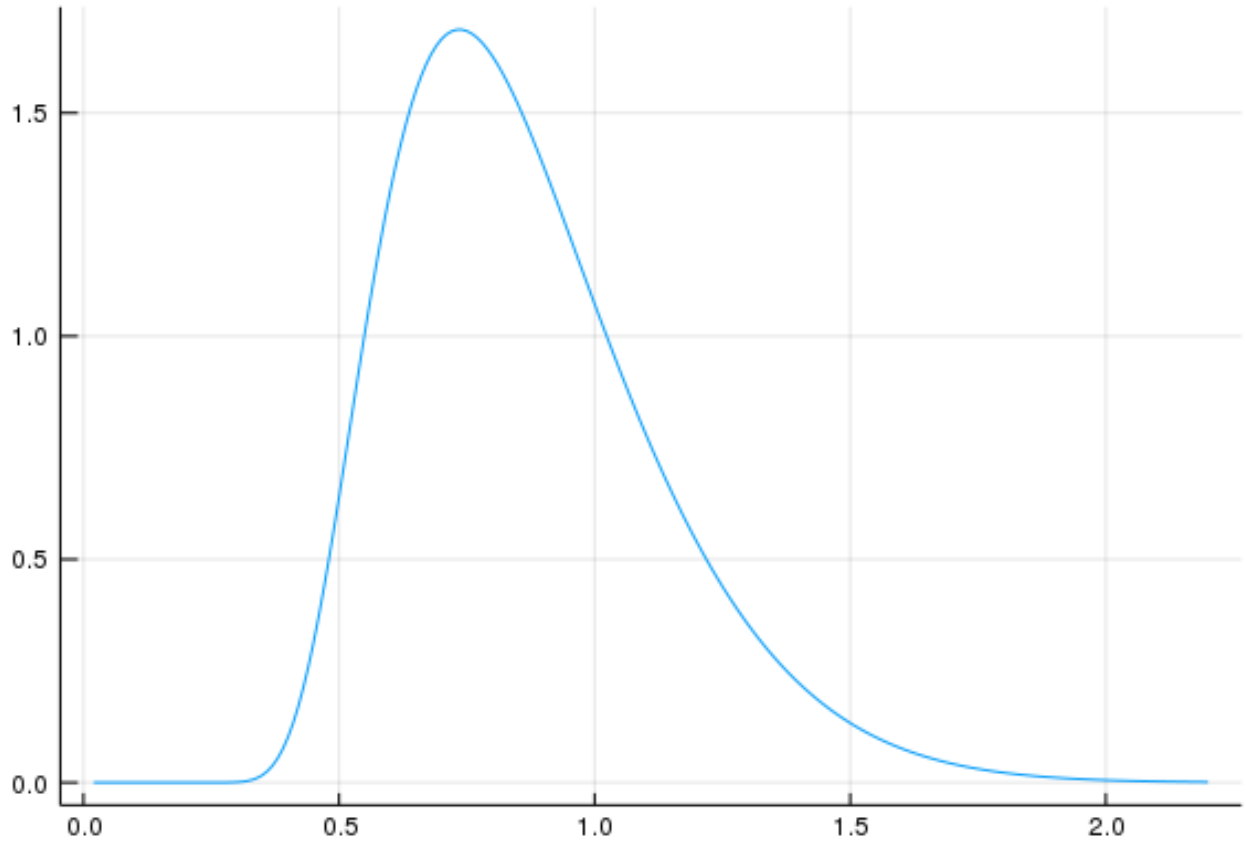


Figura: Densidad de la variable $\mathcal{K}$ de Kolmogorov

Test de hipótesis

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

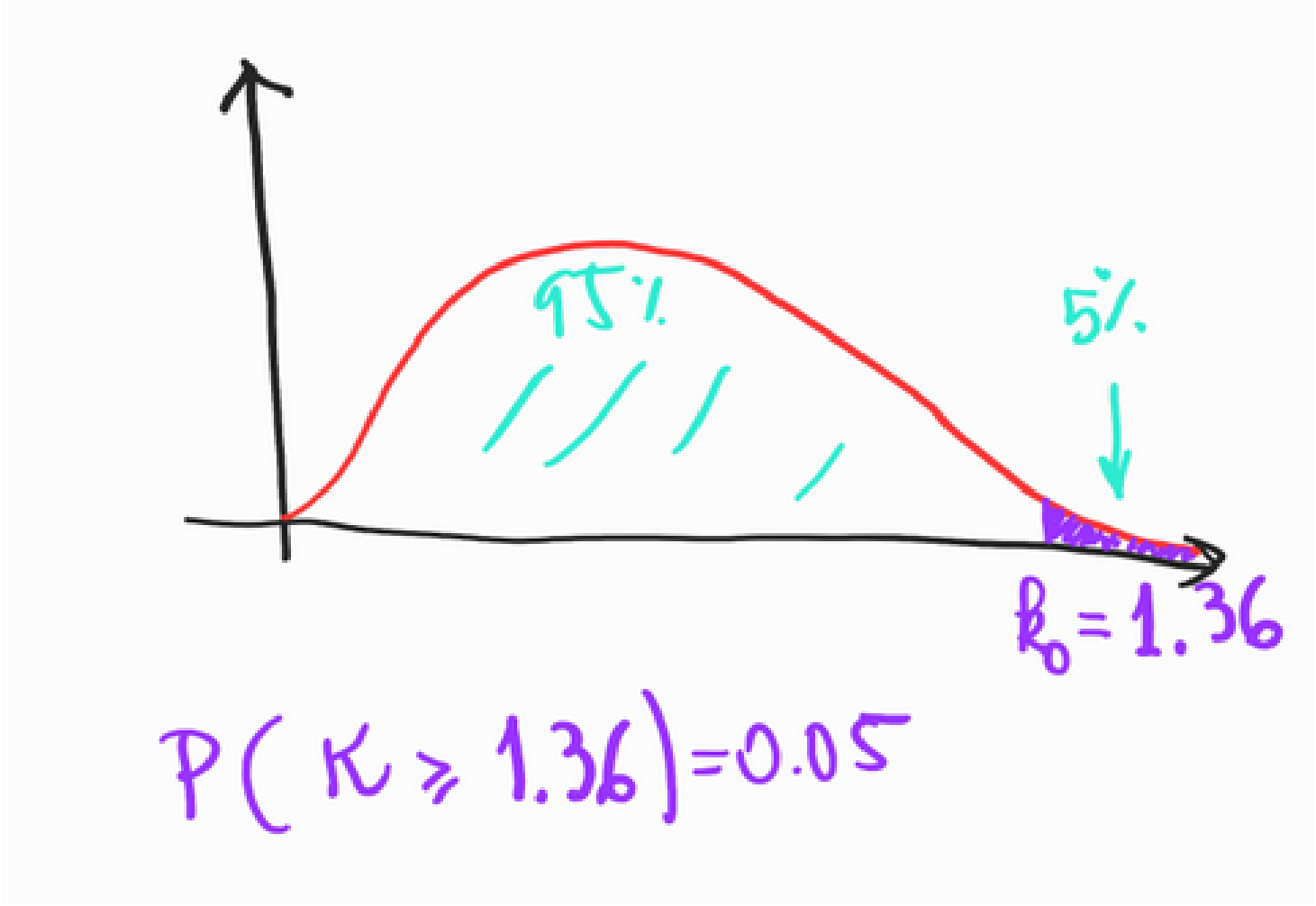
- Tenemos una muestra $X_1, \ldots, X_n$ de una distribución F y calculamos K_n.
- Construimos un intervalo de confianza para K_n de nivel $1 - \alpha$, de la forma

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

- 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:



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:

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

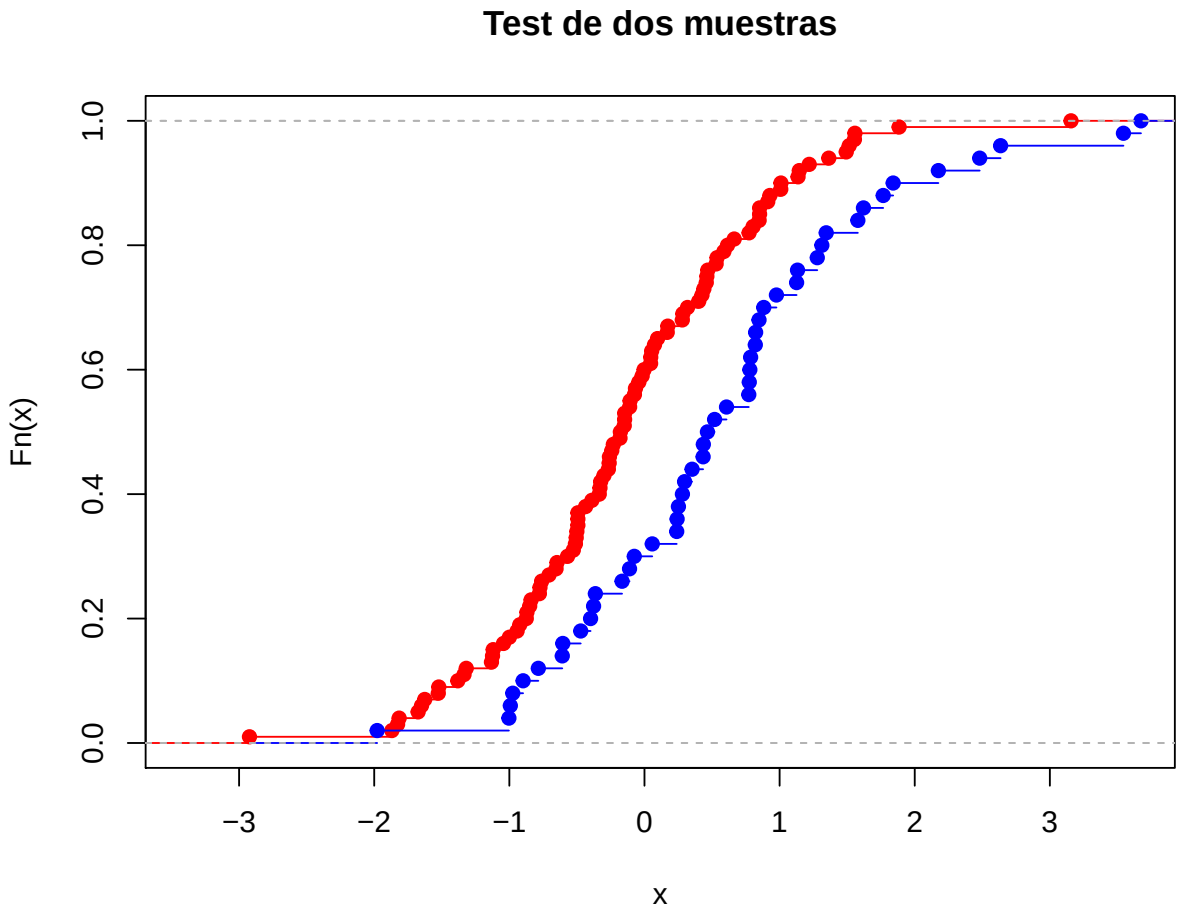
- La segunda muestra tiene m elementos

$$Y_1, \ldots, Y_m$$

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

$$H_0 : F_X = F_Y$$

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.

Test de Lilliefors

- Una situación frecuente es cuando queremos saber si los datos pertenencen a una familia de distribuciones
- Por ejemplo, si son normales (no importa cuales sean los parámetros)
- El problema que tenemos es desconocemos los parámetros de la posible distribución a testear
- ¿Qué hacemos entonces?

- estimamos los parámetros
- Medimos la distancia de Kolmogorov entre la distribución con parámetros estimados y la distribución empírica F_n
- Como ambas distribuciones dependen de los datos, ahora no se cumple el teorema de Kolmogorov
- La distribución límite ahora si depende de la familia considerada
- Lilliefors[1] calculó mediante simulación los intervalos para construir regiones críticas

[1] Hubert Whitman Lilliefors (1928 - 2008, EUA)

Test de normalidad de Lilliefors

- Queremos saber si nuestra m.a.s. $X_1, \ldots, X_n$ corresponde a alguna distribución normal

$$H_0 \colon F \in \{\mathcal{N}(\mu, \sigma^2) \colon \mu \in \mathbb{R}, \sigma > 0\}$$

- Estimamos μ y σ^2
- Tenemos la distribución empírica F_n y la distribución $\Phi(x, \hat{\mu}, \hat{\sigma}^2)$
- Calculo la distancia entre

$$D_n = \sqrt{n} \sup_x |F_n(x) - \Phi(x, \hat{\mu}, \hat{\sigma}^2)|$$

Rechazo si la distancia D_n está en la región crítica

Sample Size N	Level of Significance for $D = \text{Max} \left\| F^*(X) - S_N(X) \right\|$				
	.20	.15	.10	.05	.01
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.149	.153	.165	.180	.203
30	.131	.136	.144	.161	.187
Over 30	$\frac{.736}{\sqrt{N}}$	$\frac{.768}{\sqrt{N}}$	$\frac{.805}{\sqrt{N}}$	$\frac{.886}{\sqrt{N}}$	$\frac{1.031}{\sqrt{N}}$

Figura: Tabla tomada de *On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Lilliefors (1967) Journal of the American Statistical Association*

Test de Lilliefors en R

- Cargo mis datos en `muestra`
- Instalo el paquete `nortest`
- Cargo en R el paquete: `library(nortest)`
- Corro el comando `lillie.test(muestra)`

```
> muestra<-rnorm(100)
> lillie.test(muestra)

        Lilliefors (Kolmogorov-Smirnov) normality
        test

data:  muestra
D = 0.065636, p-value = 0.3602

> muestra2<-runif(100)
> lillie.test(muestra2)

        Lilliefors (Kolmogorov-Smirnov) normality
        test

data:  muestra2
D = 0.11872, p-value = 0.001415
```

Prueba Chi-cuadrado de Pearson

- Los métodos de testeo de distribuciones anteriores sirven para distribuciones continuas
- Si tenemos una distribución discreta utilizamos otro test, propuesto por Pearson[2]
- La idea es comparar la frecuencia observada de cada resultado con la probabilidad teórica

[2] Karl Pearson introdujo la correlación

Ejemplo: el dado equilibrado

- Queremos saber si un dado está equilibrado.
- Lo tiramos 1200 veces y obtenemos las siguientes frecuencias de los números 1 a 6.

Valor	1	2	3	4	5	6
Observado	210	180	207	212	189	202
Esperado	200	200	200	200	200	200

- ¿Cómo saber si el dado está equilibrado?

- Pearson propuso el siguiente estadístico:

$$T_n = \sum_{k=1}^{6} \frac{(e_i - o_i)^2}{e_i}$$

- Aquí
  - e_i es la cantidad esperada en la clase i
  - o_i es la cantidad observada en la clase i
- La cantidad esperada es $e_i = np_i$ donde $p_i = 1/6$ es la probabilidad de la clase

- Calculamos

$$T_n = \frac{1}{200}\Big((210-200)^2 + (180-200)^2 + (207-200)^2 + (212-200)^2 + (89-200)^2 + (202-200)^2\Big) = 4{,}09$$

- Nuestro problema es decidir si 4,09 es un desvío crítico de lo esperado, que nos haga rechazar la hipótesis nula de que el dado esté equilibrado.

- Pearson demostró un teorema central del límite, que establece que

$$T_n = \sum_{k=1}^{K} \frac{(e_i - o_i)^2}{e_i} \to \chi^2_{K-1}$$

- La variable aleatoria χ^2_ℓ (se lee Chi cuadrado con ℓ grados de libertad) es la suma de ℓ variables aleatorias normales estándar, es decir

$$\chi^2_{K-1} = (Z_1)^2 + \cdots + (Z_{K-1})^2$$

donde las Z_i son normales estándar.

- Es como que cada sumando al cuadrado converge a una normal al cuadrado, pero una de ellas desaparece.

La variable aleatoria Chi cuadrado en R

Como todas las variables tenemos cuatro comandos. El parámetro `df` son los grados de libertad[3]:

- `dchisq(x, df)` es la densidad en el punto *x*
- `pchisq(q, df)` es la distribución en el punto *x*
- `qchisq(p, df)` es el cuantil *p* (para el test)
- `rchisq(n, df)` simula *n* variables

[3] Degrees of freedom

- Teníamos $T_n = 4{,}09$, con 5 grados de libertad (uno menos que la cantidad de clases)
- Calculamos

$$\texttt{qchisq}(0{,}95, 5) = 11{,}07$$

- La región crítica con una significancia de $\alpha = 0{,}05$ es $T_n \geq 11{,}07$
- Como no estamos en la región crítica, no rechazamos H_0
- No hay evidencia estadística para suponer que el dado no esté equilibrado.

Test Chi-cuadrado. Caso general

- El caso general es análogo: La hipótesis nula es que la variable discreta toma valores en K clases con probabilidades $p_1, \ldots, p_K$
- Tenemos K clases
- Calculamos

$$T_n = \sum_{k=1}^{K} \frac{(e_i - o_i)^2}{e_i} \to \chi^2_{K-1}$$

- Calculamos el cuantil α mediante `qchisq(`α`,K-1)`
- Si el estadístico T_n es mayor que el cuantil, rechazamos la hipótesis nula
- En caso contrario no hay evidencia estadística para rechazar la hipótesis nula