
¡TODO EL DÍA PENSANDO EN NÚMEROS, CIFRAS, CUENTAS, CÁLCULOS MATEMÁTICOS, PORCENTAJES! ¿NO TE DAS CUENTA DE CUÁNTO DAÑA NUESTRA RELACIÓN?
SÍ, UN 63%...
KAFFEL

Clase 15 de Bioestadística

Estadística: distribución empírica y cuantiles

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Probabilidad y Estadística (repaso)

Estimación de una probabilidad (repaso)

Estimación de la función de distribución

Probabilidad y Estadística (repaso)

- La probabilidad y la estadística son disciplinas hermanas
- Comparten un mismo marco teórico matemático:
  - un espacio de sucesos Ω
  - un modelo de probabilidad P
- Se diferencian en las preguntas que formula cada una:
  - La probabilidad conoce P y calcula a partir de ella
  - La estadística conoce valores $X_1, \ldots, X_n$ y se ocupa de estimar[1] P.
  - La estimación de la probabilidad se designa $\widehat{P}$.

[1] Se trata de dar una buena aproximación del modelo

Terminología estadísitica

- El conunto Ω se llama espacio muestral. En probablidad se llama espacio de sucesos
- Los valores $X_1, \ldots, X_n$ se llaman muestra aleatoria simple, se resume m.a.s[2]
- Si el parámetro desconocido se designa θ, la aproximación que hacemos se designa $\widehat{\theta}$.

[2] Es lo que en probabilidad se llama v.a.i.i.d.

Estimación de la esperanza

- La base de la estimación es la ley fuerte de los grandes números (LFGN):
- Sean $X_1, \ldots, X_n$ una m.a.s. de v.a. con esperanza desconocida $\mu = E(X_1)$.
- La LFGN establece que:

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \to E(X_1)$$

- Si tenemos un modelo con una probabilidad P con $\mu = E(X_1)$ desconocido, podemos estimar μ a traves del promedio de los datos

$$\widehat{\mu} = \overline{X}_n$$

- La LFGN nos dice entonces que

$$\widehat{\mu} \to \mu,$$

que se denomina consistencia del estimador.

Estimación de una probabilidad (repaso)

- Supongamos un modelo de Bernoulli con probabilidad p desconocida.
- Tenemos una m.a.s. $X_1, \ldots, X_n$. Se verifica

$$E(X_1) = p$$

- Entonces se reduce al caso anterior: ¡la probabilidad es la esperanza!

- Entonces, el estimador de la probabilidad es

$$\widehat{p} = \frac{X_1 + \cdots + X_n}{n} = \frac{\text{cantidad de éxitos}}{n}$$

- El estimador también se denomina frecuencia observada de la cantidad de éxitos

- Además, la LFGN nos dice que el estimador es consistente, es decir

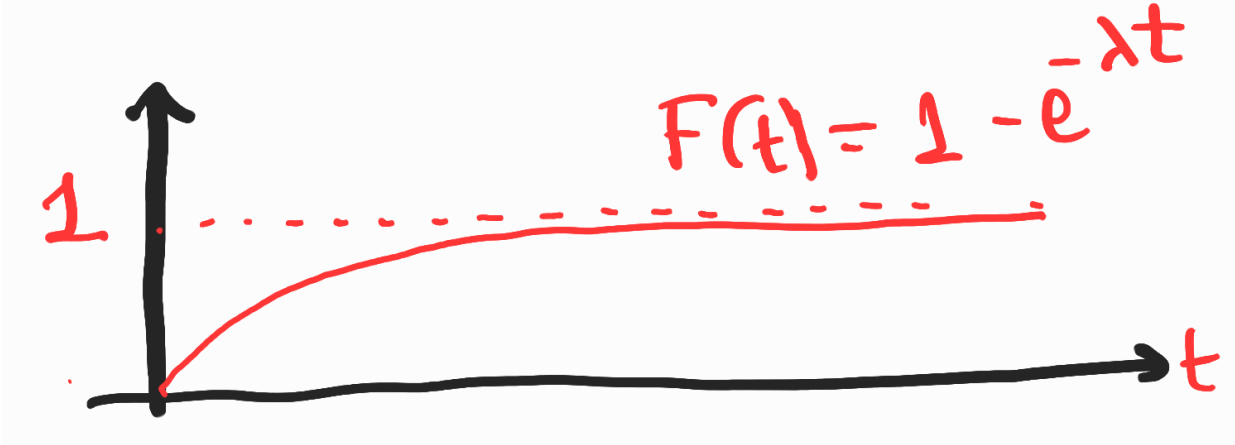
$$\widehat{p} \to p.$$

Estimación de la función de distribución

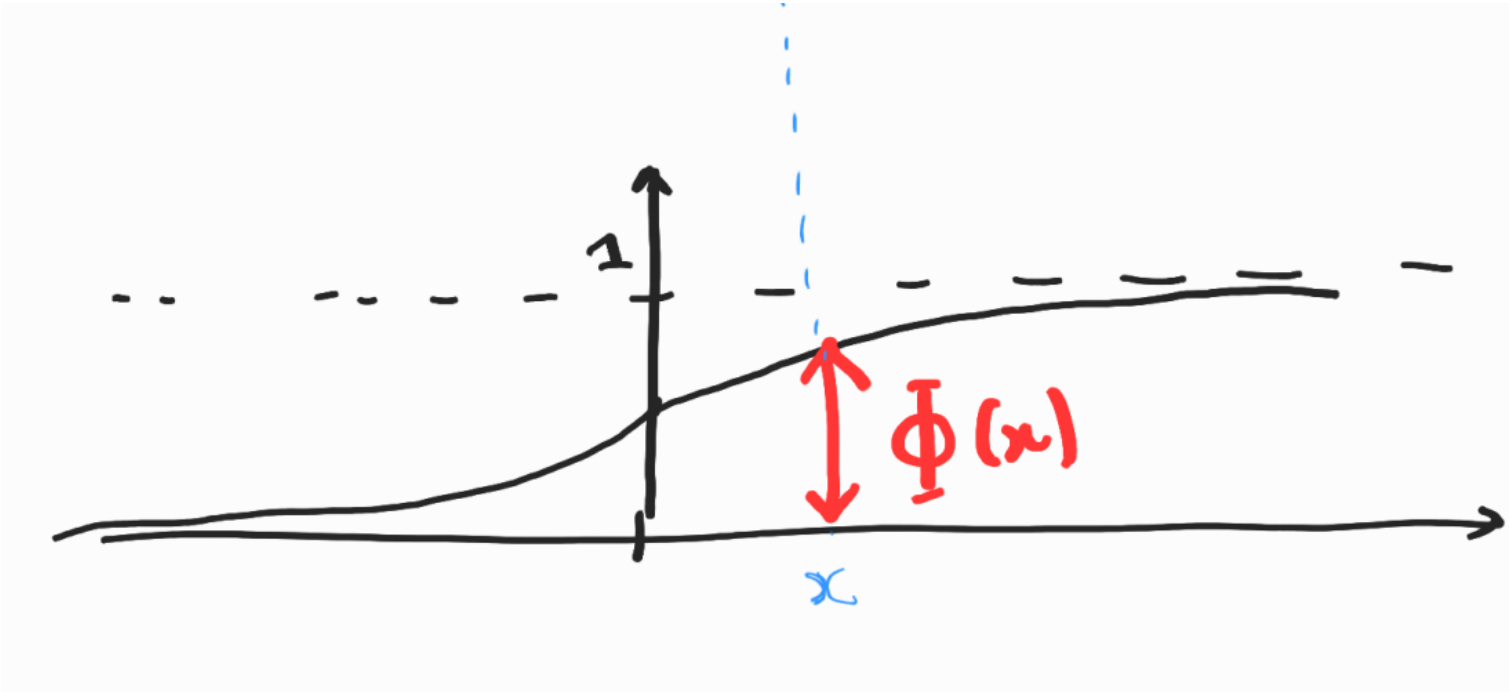
- Dada una variable aleatoria X, definimos su distribución F como la función tal que

$$F(x) = P(X \leq x)$$

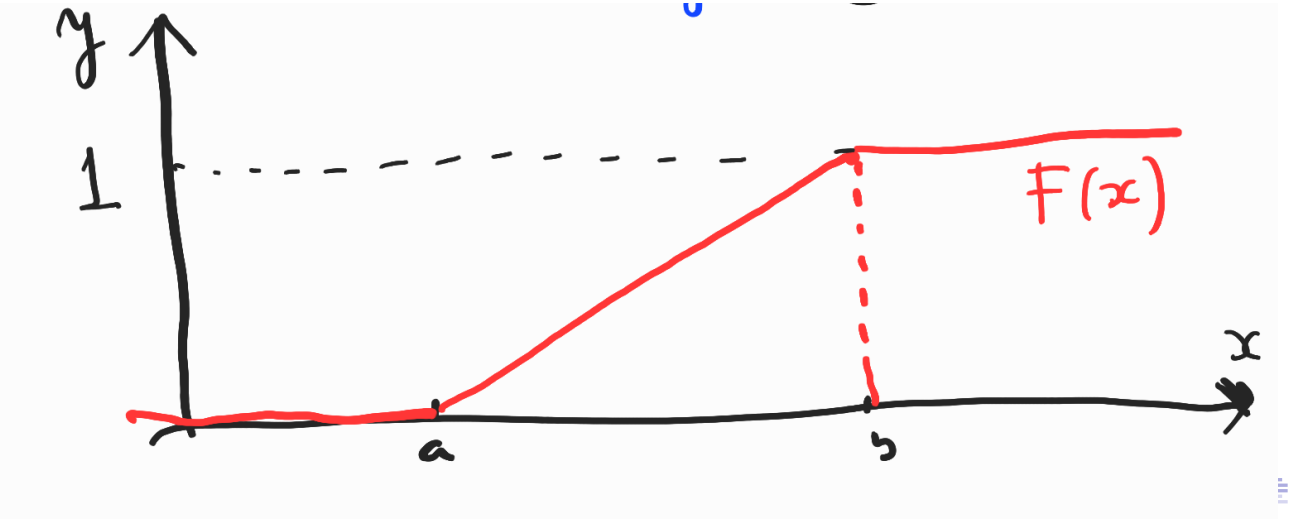
- Distribución exponencial:



- Distribución Normal:



- Distribución uniforme:



Propiedades de la función de distribución

- La distribución $F(x)$ es la probabilidad del suceso

$$\{X \leq x\},$$

por lo tanto

$$0 \leq F(x) \leq 1, \qquad \text{para todo } x \in \mathbb{R}$$

- Monotonía: Si $a < b$ tenemos la inclusión de sucesos

$$\{X \leq a\} \leq \{X \leq b\},$$

(si X (el resultado del experimento) es menor o igual que a y a es menor que b entonces el resultado X es menor o igual que b). Eso nos da

$$F(a) \leq F(b)$$

- Si $x = -\infty$ entonces $F(x) = 0$: nunca ocurre $X \leq -\infty$, entonces

$$\{X \leq -\infty\} = \emptyset, \quad \text{luego} \quad F(-\infty) = P(\emptyset) = 0$$

- Si $x = +\infty$ entonces $F(x) = 1$: siempre ocurre $X \leq +\infty$, entonces

$$\{X \leq +\infty\} = \Omega, \quad \text{luego} \quad F(+\infty) = P(\Omega) = 1.$$

Estimación de la función de distribución

Nos planteamos el siguiente problema:

- Tenemos una muestra aleatoria simple

  $$X_1, \ldots, X_n$$

  que corresponde a una v.a. con distribución F
- Desconocemos F y queremos estimarla
- F es una función, empecemos por un valor $F(a)$.

- $F(a) = p$ es la probabilidad del suceso

$$\{X \leq a\}$$

- El estimador de una probabilidad p es su frecuencia, es decir, la proporción de casos favorables
- Ejemplo: Supongamos que $a = 0$ y nuestra muestra son los valores

$$0{,}55,\ 0{,}91,\ -0{,}16,\ -0{,}30,\ -0{,}83,\ -0{,}59,\ 0{,}45,\ -0{,}29$$

- ¿Cuáles son nuestros éxitos? Como $a = 0$ el suceso es

$$X \leq 0$$

es decir, son los valores negativos:

$$0{,}55,\ 0{,}91,\ -0{,}16,\ -0{,}30,\ -0{,}83,\ -0{,}59,\ 0{,}45,\ -0{,}29$$

- Entonces nuestra estimación es

$$\widehat{p} = \frac{5}{8}.$$

- Para el estimador de $F(a)$ se usa la notación $F_n(a)$, que se llama distribución empírica de F. Tenemos entonces

$$F_n(0) = \frac{5}{8}$$

- Hacemos el mismo razonmiento ahora para cualquier a:
- Para estimar $F(a)$ contamos la cantidad de variables de la muestra que son menores o iguales que a y dividimos por el tamaño de la muestra, es decir

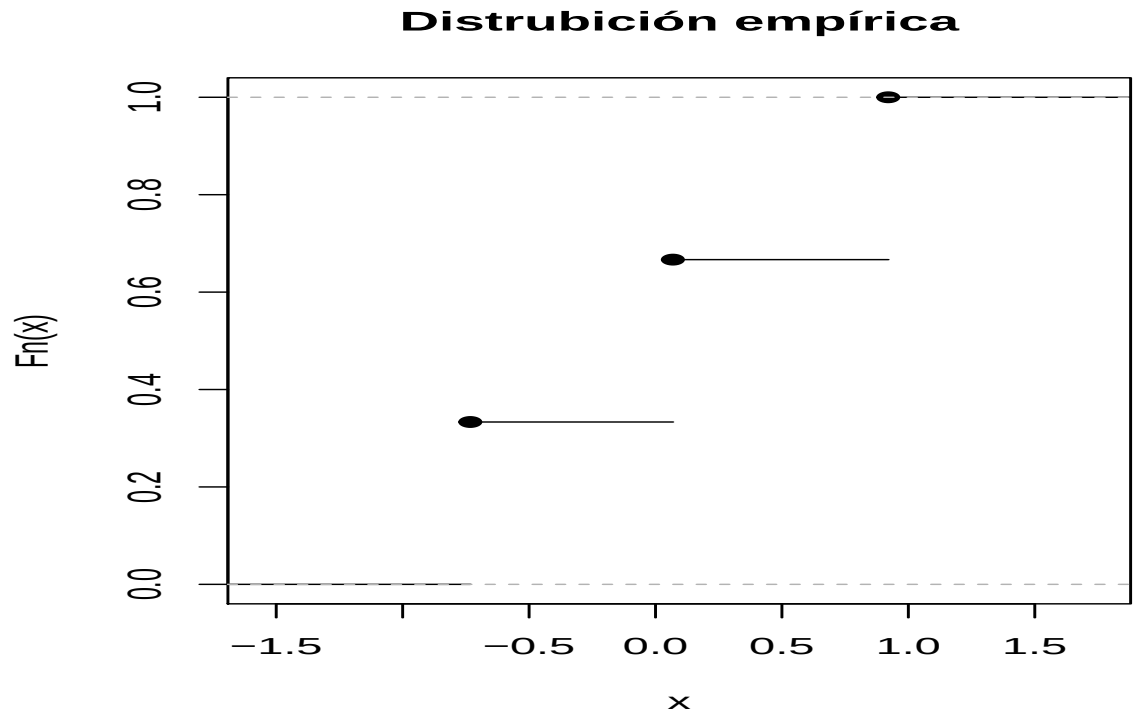
$$F_n(a) = \frac{\#\{k : X_k \leq a\}}{n}$$

- Miramos ahora la función $F_n(a)$ cuando varía a, si por ejemplo $n = 3$ y tenemos los valores

$$0{,}07,\ 0{,}92,\ -0{,}73$$

si $a < -0{,}73$ tenemos $F_3(a) = 0$.

- Si $-0{,}73 \leq a < 0{,}07$ tenemos $F_3(a) = 1/3$
- Si $0{,}07 \leq a < 0{,}92$ tenemos $F_3(a) = 2/3$
- Si $0{,}98 \leq a$ tenemos $F_3(a) = 1$



Para plotear la distribución empírica en R se usa el comando `ecdf`. Concretamente el código es

```
muestra<-rnorm(3,0,1)
p<-ecdf(muestra)
plot(p,main="Distribución empírica")
```

Consistencia

Ahora, ¿se parece la F_n (estimada) la F (teórica)?

- Para eso tenemos la indicatriz de un conjunto:

$$\mathbf{1}_A = \begin{cases} 1, & \text{si } x \in A \\ 0, & \text{si } x \notin A \end{cases}$$

- Es decir, sumamos

$$\#\{k : X_k \leq a\} = \sum_{k=1}^{n} \mathbf{1}_{\{X_k \leq a\}}$$

- Entonces tenemos un promedio, y aplicamos la LFGN:

$$F_n(a) = \widehat{p} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{X_k \leq a\}} \rightarrow E\left(\mathbf{1}_{\{X_1 \leq a\}}\right)$$

- Como

$$E\left(\mathbf{1}_{\{X_1 \leq a\}}\right) = 0 \times P(X_1 > a) + 1 \times P(X_1 \leq a) = P(X_1 \leq a) = F(a)$$

obtenemos la consistencia de la estimación:

$$F_n(a) \rightarrow F(a)$$

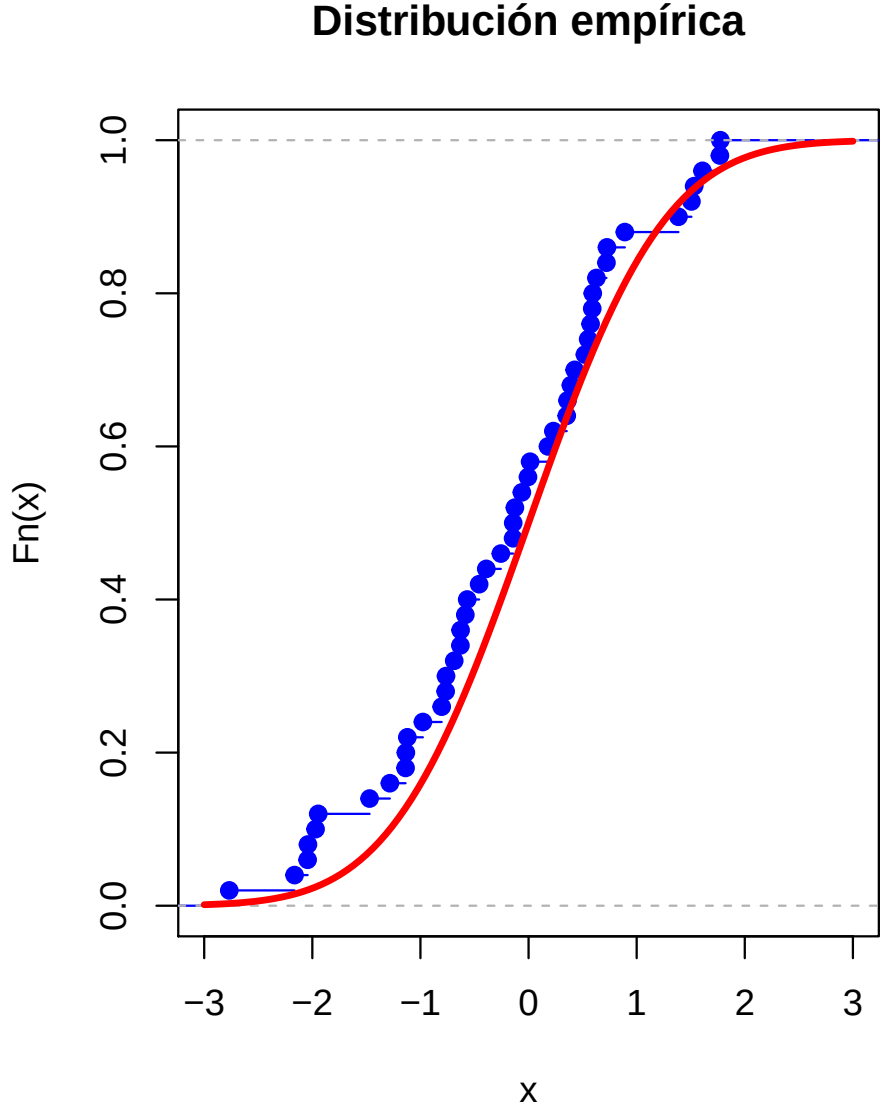


Figura: La distribución empírica (azul) se aproxima a la distribución teórica (roja)