

1. Espacios de Probabilidad

Probabilidades y Estadística (M)

María Eugenia Szretter Noste

Departamento de Matemática e
Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Primer cuatrimestre 2020



Experimentos

Llamamos **experimento** a un proceso que genera un resultado. Un experimento puede ser

- **determinista** (las condiciones en las que se realiza el experimento determinan el resultado), como el *MRU movimiento rectilíneo uniforme*. La posición inicial y la velocidad determinan la posición del objeto en cada instante,

$$\underbrace{p(t)}_{\text{posición a tiempo } t} = \underbrace{p_0}_{\text{posición inicial}} + \underbrace{v}_{\text{velocidad}} \underbrace{t}_{\text{tiempo}}$$

- **aleatorio** (no podemos prever el resultado del mismo), como el *paseo al azar unidimensional*.

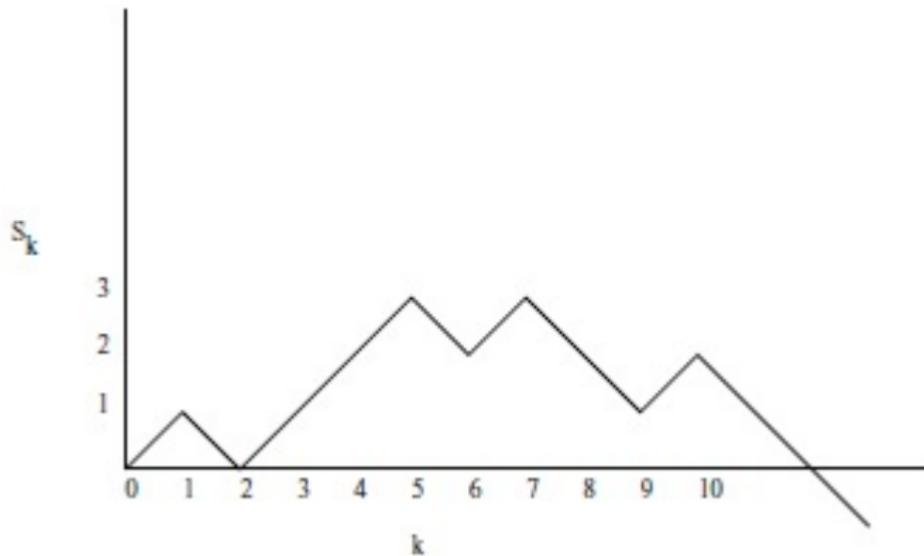
Paseo al azar unidimensional

Una partícula se desplaza en la recta, comenzando en el origen de \mathbb{Z} . Realiza n pasos aleatorios de la siguiente manera. En cada momento $t = 1, 2, \dots, n$ lanzamos una moneda. Si sale cara, la partícula se mueve una unidad hacia la derecha, si sale ceca, la partícula se mueve una unidad hacia la izquierda. ¿Dónde se encuentra la partícula a tiempo n ? Podemos indicar a qué puntos podría haber llegado y a cuáles no, pero no podemos decir dónde se encuentra. Podemos anotar el **espacio de posibles resultados del experimento aleatorio**: es $\{-1, +1\}^n$, el conjunto de secuencias de longitud n con los símbolos -1 y $+1$. Por ejemplo, si $n = 4$ y $\omega = (+1, -1, +1, +1)$, entonces el primer paso fue a la derecha, el segundo a la izquierda, el tercero y el cuarto a la derecha nuevamente y la partícula se encuentra en la posición

$$S_k(\omega) = \sum_{i=1}^k \omega_i$$

para $k = 4$. Podemos asociarle a este experimento el siguiente gráfico

Paseo al azar unidimensional

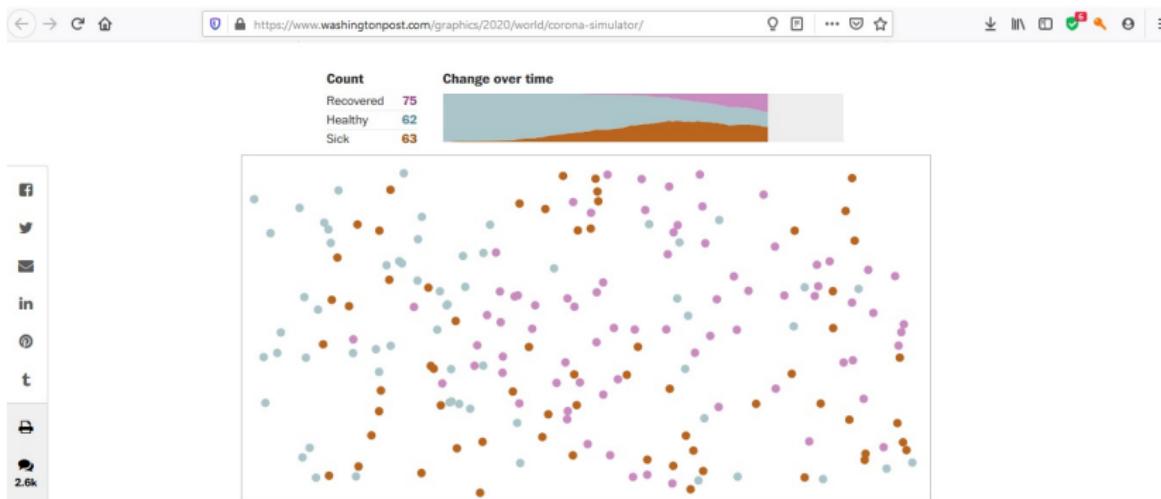


En esta materia nos interesan los experimentos aleatorios.

Un experimento más: nota del Washington Post

Un ejemplo de mucha relevancia actual son las simulaciones de contagio de enfermedad. En el link que sigue hay una versión simplificada del impacto del aislamiento social en el contagio

<https://www.washingtonpost.com/graphics/2020/world/corona-simulator/>



Nota del Washington Post sobre brotes epidémicos

Este experimento aleatorio consiste en ubicar 200 individuos (representados por bolitas) al azar en un rectángulo. Estos individuos se mueven al azar (como si se desplazaran por la ciudad donde viven). Uno de ellos está enfermos (**color marrón**) al inicio del experimento. Cuando toca a otro individuo sano (**celeste**) lo contagia. Para cada individuo, cuando pasa un determinado tiempo fijo desde que se enfermó, se sana y cambia al color **violeta**. Mientras la simulación avanza se lleva la cuenta a cada instante de tiempo de cuantos individuos de cada categoría hay (es el gráfico en la parte superior). Se simulan 4 escenarios:

- ① movimiento usual
- ② cuarentena forzada de enfermos (imperfecta)
- ③ “distanciamiento social” moderado
- ④ “distanciamiento social” estricto

y se comparan los números de infectados a tiempo t para cada escenario. Observar la palabra “**simulación**” asociada a cada repetición del experimento.

Espacio muestral

Definicion 1.1

Dado un experimento, llamamos *espacio muestral* al conjunto Ω formado por todos los posibles resultados del experimento. También utilizaremos \mathcal{S} para denotar el espacio muestral.

Ejemplo 1.1

1. Si el experimento consiste en lanzar un dado, tenemos
$$\Omega_1 = \{1, 2, 3, 4, 5, 6\}.$$
 Observemos que el espacio muestral es finito y consta de 6 elementos.
2. Si lanzamos un dado rojo y otro negro, el espacio muestral es

$$\Omega_2 = \{(x_1, x_2) : x_i \in \mathbb{N}, 1 \leq x_1, x_2 \leq 6\}$$

donde la primer componente indica el resultado del dado rojo y la segunda componente indica el resultado del dado negro. El espacio muestral es finito y consta de 36 elementos.

Ejemplo 1.1 cont., espacios muestrales

3. Contamos el número de lanzamientos de un dado hasta la primera vez que observamos un seis, tenemos que

$$\Omega_3 = \{1, 2, 3, \dots\}.$$

A diferencia del caso anterior, Ω_3 es un conjunto infinito (numerable).

4. Si el experimento consiste en lanzar un dardo sobre un tablero cuadrado de dos por dos, el espacio muestral estará formado por todos los posibles puntos del cuadrado:

$$\Omega_4 = \{(x, y) \in \mathbb{R}^2 : 0 \leq x, y \leq 2\}.$$

A diferencia de los ejemplos anteriores, este espacio muestral está formado por una cantidad no numerable de puntos.

5. Observamos el tiempo de duración de una lamparilla. En este caso tenemos que el espacio muestral está formado por todos los números reales positivos: $\Omega_5 = [0, +\infty)$.

Experimentos aleatorios (en R)

www.r-project.org

Un muy buen video introductorio

<https://www.youtube.com/watch?v=5UDTAnBhF9E&feature=youtu.be>

```
# 1. un dado
```

```
sample(1:6,1)
```

```
# 2. dos dados: rojo y negro
```

```
sample(1:6,2,replace=T)
```

```
# 3. un dado hasta obtener el primer seis
```

```
a<-0
```

```
tiros<-0
```

```
while (a<6){
```

```
  a<-sample(1:6,1)
```

```
  tiros<-c(tiros,a)}
```

```
  tirostot<-tiros[-1]
```

```
  numtiros<-length(tirostot)
```

Experimentos aleatorios (en R), continuación

```
# 4. un punto en el tablero de 2 x 2
x1<-runif(1,min=0,max=2)
x2<-runif(1,min=0,max=2)
c(x1,x2)
plot(x1,x2,xlim = c(0,2),ylim=c(0,2),pch=15,col="red")

#Repetimos 10 veces (tiramos 10 dardos)
x1<-runif(10,min=0,max=2)
x2<-runif(10,min=0,max=2)
c(x1,x2)
plot(x1,x2,xlim = c(0,2),ylim=c(0,2),pch=15,col="red")

# 5. duracion de la lamparita
rexp(1,rate=1/100)
# lo repetimos 15 veces
rexp(15,rate=1/100)
```

Eventos

Tras haber identificado el espacio muestral asociado a nuestro experimento, estaremos interesados en identificar posibles subconjuntos de éste.

Definicion 1.2

Un *evento (o suceso)* es un conjunto formado por algunos de los resultados posibles del experimento. En otras palabras, un evento es un subconjunto A incluido en el espacio muestral Ω . Generalmente, utilizaremos letras mayúsculas (A, B, C) para denotar eventos.

Definicion 1.3

Llamamos *evento elemental* a aquellos constituidos por un único elemento. Es decir, A se dice elemental si

$$A = \{a\}, \text{ para algún } a \in \Omega .$$

Ejemplos de eventos

Consideremos algunos eventos relacionados con los experimentos dados:

1. Cuando lanzamos el dado, podríamos considerar el suceso A : *el resultado del lanzamiento es par*, teniendo que

$$A = \{2, 4, 6\} .$$

2. Para el ejemplo de los dos dados, podríamos estar interesado en el evento A , *el resultado del dado rojo es un uno*, obteniendo así que

$$A = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)\} .$$

Si estuviéramos interesados en el evento B : *la suma de los dados da 7*, tendríamos que

$$B = \{(1, 6), (2, 5), (3, 4), (5, 2), (6, 1)\} .$$

Ejemplos de eventos

3. Al contar el número de lanzamientos hasta obtener el primer seis, podríamos considerar el evento B : *tardamos a lo sumo 8 tiros*, en cuyo caso tendríamos que

$$B = \{1, 2, 3, 4, 5, 6, 7, 8\}.$$

Si el evento C consiste en *tardamos una cantidad impar de tiros hasta obtener el primer seis*, tendríamos que

$$C = \{1, 3, 5, 7, 9, \dots\}.$$

También podríamos considerar el evento D donde *tardamos al menos 15 tiros en observar el primer 6*, obteniendo que

$$D = \{15, 16, 17, \dots\}.$$

4. Tirando el dardo, podemos considerar el evento A : *el dardo cayó en la mitad superior*, teniendo así que

$$A = \{(x, y) \in \Omega : y \geq 1\}.$$

Operaciones con conjuntos

Dados dos eventos A y B , tenemos que

$$A \cup B = \{x : x \in A \text{ ó } x \in B\}.$$

Es decir: la unión de dos eventos está constituida por los elementos que están en uno u otro conjunto. En general, nuestro “o” coloquial se transforma en una unión al operar con eventos.

Ejemplo: Al contar el número de lanzamientos del dado, podríamos estar interesados en que éste sea a lo sumo ocho o en que el número de lanzamientos sea par.

¿Qué acontece si queremos que el número de lanzamientos sea par y a lo sumo ocho? En este caso, estamos imponiendo dos restricciones. Esta operación se interpreta matemáticamente como la intersección de los eventos.

$$A \cap B = \{x : x \in A \text{ y } x \in B\}.$$

Operaciones con conjuntos

Podemos estar interesados en que no ocurra cierto evento A . De esta forma obtendremos el complemento de A :

$$A^c = \{x : x \notin A\}.$$

Finalmente, definimos la diferencia entre eventos $A - B$ de la siguiente forma:

$$A - B = A \cap B^c = \{x : x \in A \text{ y } x \notin B\}.$$

Propiedades de operaciones con conjuntos

Lema 1.1

Para las operaciones de conjuntos definidas, vale que

1. *Asociatividad:*

$$(A \cup B) \cup C = A \cup (B \cup C), \quad (A \cap B) \cap C = A \cap (B \cap C).$$

2. *Distributividad:*

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

3. *Leyes de De Morgan:*

$$\left(\bigcup_{\alpha} A_{\alpha} \right)^c = \bigcap_{\alpha} A_{\alpha}^c, \quad \left(\bigcap_{\alpha} A_{\alpha} \right)^c = \bigcup_{\alpha} A_{\alpha}^c.$$

4. *Finalmente, notemos que*

$$(A^c)^c = A, \quad A \cup \emptyset = A, \quad A \cap \emptyset = \emptyset, \quad A \cap \Omega = A.$$

Definicion 1.4

Dos eventos A y B se dicen *disjuntos* o *excluyentes* si

$$A \cap B = \emptyset .$$

Idea Intuitiva de la Probabilidad

Nuestro próximo objetivo es asignar a cada evento A un número $P(A)$, con $0 \leq P(A) \leq 1$, que indique la chance que tenemos de que el resultado del experimento pertenezca al evento A . Repetimos el experimento, para lograrlo. Sean

m = número de veces que repetimos el experimento

m_A = número de veces que el resultado del experimento pertenece al evento A en las m repeticiones

$\frac{m_A}{m}$ = frecuencia relativa del evento A en m repeticiones

Es de esperar que a medida que el número de repeticiones crece, **la frecuencia relativa se estabiliza**, convergiendo a cierto valor que es lo que interpretaremos como la probabilidad de A , $P(A)$.

Intuitivamente, entonces

$$P(A) := \lim_{m \rightarrow \infty} \frac{m_A}{m}.$$

Teoría axiomática de la probabilidad

La teoría de probabilidades procura dar rigor a la idea previa. Tiene 3 ingredientes.

- ① Ω (espacio muestral): conjunto de posibles resultados del experimento.
- ② Clase de conjuntos a los cuales les vamos a definir una función de probabilidad.
- ③ Función de probabilidad.

Definición 1.5

Dado Ω , consideramos

$$\mathcal{P}(\Omega) = \{A : A \subseteq \Omega\}$$

la familia de *partes* de Ω . Notemos que si Ω es finito y consta de N elementos, entonces $\mathcal{P}(\Omega)$ también es finito y tiene 2^N elementos .

Clase de conjuntos

¿A qué subconjuntos de Ω podemos definirles una probabilidad, (de modo que se cumplan algunas propiedades básicas que cumple la frecuencia relativa)? Resulta que no siempre vamos a poder asignar probabilidad a todo subconjunto de cualquier Ω (es un resultado que se probará en *Análisis Real*). Entonces buscamos una subcolección que nos interese.

Definición 1.6

Dado una espacio muestral Ω , un *álgebra* \mathcal{A} en Ω es una familia formada por eventos de Ω ($\mathcal{A} \subset \mathcal{P}(\Omega)$), verificando las siguientes propiedades:

- ① $\Omega \in \mathcal{A}$.
- ② Si $A \in \mathcal{A}$, entonces, $A^c \in \mathcal{A}$.
- ③ Si $A, B \in \mathcal{A}$, entonces $A \cup B \in \mathcal{A}$.

Observación 1

Una unión finita de elementos de \mathcal{A} también estará en \mathcal{A} . ¿Nos basta con uniones finitas?

¿Cuánto debemos agrandar la colección de conjuntos?

¿Por qué queríamos incluir las uniones numerables entre los eventos a los que queremos poder asignarles probabilidades?

Ejemplo 1.2

Al contar el número de lanzamientos de un dado hasta obtener el primer seis, claramente debemos ser capaces de calcular las probabilidades de los eventos elementales $\{1\}$, $\{2\}$, También queremos ser capaces de calcular la probabilidad del evento *C tardamos una cantidad impar de tiros hasta obtener el primer seis*, tendríamos que

$$C = \{1, 3, 5, 7, 9, \dots\} = \bigcup_{i=1}^{\infty} \{2i - 1\},$$

que es una **unión numerable de eventos**.

Sin embargo, un álgebra no es necesariamente cerrada bajo uniones numerables, como lo muestra el siguiente ejemplo.

Ejemplo 1.3

Consideremos en $\Omega = \mathbb{R}$, la clase \mathcal{A} formada por uniones disjuntas de finitos intervalos (uniones finitas de intervalos disjuntos entre sí). Es decir,

$$\mathcal{A} = \left\{ A = \bigcup_{i=1}^n I_i, n \geq 0 \quad I_i \cap I_j = \emptyset \text{ para } i \neq j \right\}$$

donde I_i puede tomar la forma $(-\infty, a]$, $(a, b]$, $(b, +\infty)$, $(-\infty, +\infty)$ y para $n = 0$ la unión vacía es el propio conjunto vacío. Es fácil verificar que \mathcal{A} es, efectivamente, un álgebra. Consideremos los conjuntos

$A_i = (0, 1 - \frac{1}{i}] \in \mathcal{A}$, ($i \in \mathbb{N}$, $i \geq 2$). Sin embargo

$$\bigcup_{i \geq 2} A_i = (0, 1) \notin \mathcal{A}.$$

Por ello, agrandamos las familias de subconjuntos del espacio muestral que nos interesan.

Definición 1.7

Dado un espacio muestral Ω , una σ -álgebra en Ω es una familia \mathcal{F} formada por eventos de Ω , verificando las siguientes propiedades:

- ① $\Omega \in \mathcal{F}$.
- ② Si $A \in \mathcal{F}$, entonces, $A^c \in \mathcal{F}$.
- ③ Si $A_i \in \mathcal{F}$ para $i \geq 1$, entonces $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

En inglés se las denomina σ -fields.

También son cerradas bajo intersecciones numerables (ejercicio).

Función de probabilidad

Definición 1.8

Una *probabilidad* definida en el par (Ω, \mathcal{F}) es una función

$$P: \mathcal{F} \rightarrow [0, 1]$$

verificando las siguientes propiedades:

- ① La probabilidad del espacio muestral es uno: $P(\Omega) = 1$.
- ② σ -aditividad: Si tenemos una sucesión de eventos $(A_i)_{i \geq 1}$, $A_i \in \mathcal{F}$ para $i \geq 1$, disjuntos dos a dos (es decir, $A_i \cap A_j = \emptyset$ para $i \neq j$), entonces

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

La terna (Ω, \mathcal{F}, P) constituida por un conjunto Ω , una σ -álgebra \mathcal{F} y una probabilidad P definida en (Ω, \mathcal{F}) se llama *espacio de probabilidad*.

Lema 1.2

Toda función de probabilidad verifica las siguientes propiedades:

- 1) $P(\emptyset) = 0$. Dem: Como el \emptyset es disjunto consigo mismo, $\emptyset \cap \emptyset = \emptyset$, la sucesión $\emptyset, \emptyset, \emptyset, \dots$ es disjunta. Luego

$$P(\emptyset) = P(\emptyset \cup \emptyset \cup \emptyset \cup \dots) = \sum_{i=1}^{\infty} P(\emptyset)$$

pero esta igualdad sólo puede ser válida si $P(\emptyset) = 0$.

- 2) Aditividad finita: Si $(A_i)_{1 \leq i \leq n}$ son eventos disjuntos dos a dos ($A_i \cap A_j = \emptyset$ para $i \neq j$), entonces $P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$.

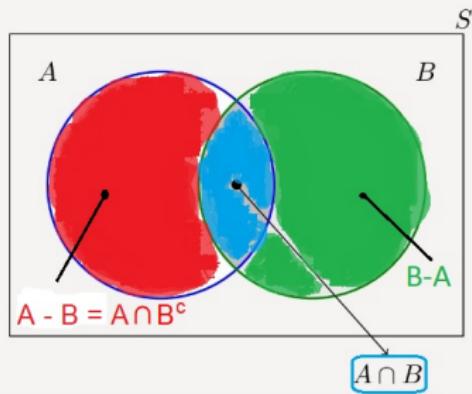
Dem: Definimos $A_i = \emptyset \ \forall i > n$, entonces $P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \underbrace{\sum_{i=1}^{\infty} P(A_i)}_{\text{por Lema 1.2 1}} = \sum_{i=1}^n P(A_i)$.

- 3) $P(A^c) = 1 - P(A)$. Dem: Caso particular de 2) tomando A y A^c .

Lema 1.2, continuación

4) Dados dos eventos A y B , tenemos que

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$



Observemos que valen las uniones disjuntas

$$A = (A \cap B^c) \cup (A \cap B)$$

$$B = (A \cap B) \cup (A^c \cap B)$$

$$A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$$

Por la aditividad finita de P , resulta

$$P(A) = P(A \cap B^c) + P(A \cap B) \quad (1)$$

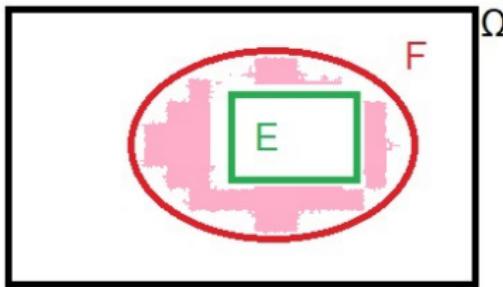
$$P(B) = P(A \cap B) + P(A^c \cap B) \quad (2)$$

y también

$$\begin{aligned} P(A \cup B) &= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) \\ &\stackrel{(1)}{=} P(A) + P(A^c \cap B) \stackrel{(2)}{=} P(A) + P(B) - P(A \cap B) \end{aligned}$$

Lema 1.2, continuación

5) Monotonía: Si $E \subset F$, entonces $P(F) = P(F - E) + P(E)$; en particular, tenemos que $P(E) \leq P(F)$.



Demostración.

Descomponemos $F = E \cup (F - E)$ (unión disjunta) y usamos Lema 1.2 2) (aditividad finita): $P(F) = P(E) + P(F - E)$. □

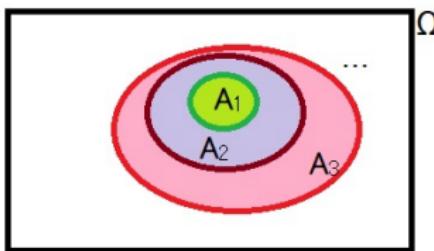
Lema 1.2, continuación

- 6) *Continuidad de la probabilidad: Dada una sucesión creciente de eventos $A_1 \subseteq A_2 \subseteq A_3 \dots$, entonces*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \rightarrow \infty} P(A_n).$$

Si tenemos eventos decrecientes $E_1 \supseteq E_2 \supseteq E_3 \supseteq \dots$, entonces

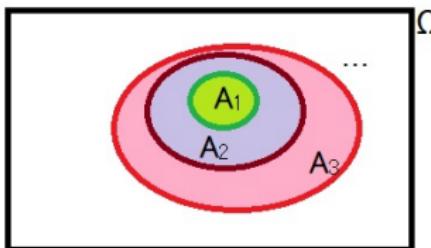
$$P\left(\bigcap_{i=1}^{\infty} E_i\right) = \lim_{n \rightarrow \infty} P(E_n).$$



Dem: (consideraremos los “anillos” **disjuntos** pintados en el dibujo)

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_2 - A_1) \cup (A_3 - A_2) \cup \dots$$

$$= A_1 \cup \bigcup_{i=1}^{\infty} (A_{i+1} - A_i)$$



$$\begin{aligned} \bigcup_{i=1}^{\infty} A_i &= A_1 \cup (A_2 - A_1) \cup (A_3 - A_2) \cup \dots \\ &= A_1 \cup \bigcup_{i=1}^{\infty} (A_{i+1} - A_i) \quad (\text{disj}) \end{aligned}$$

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} A_i\right) &= P(A_1) + \sum_{i=1}^{\infty} P(A_{i+1} - A_i) \\ &= P(A_1) + \sum_{i=1}^{\infty} [P(A_{i+1}) - P(A_i)] \quad (\text{prop 5, } A_i \subset A_{i+1}) \\ &= P(A_1) + \lim_{n \rightarrow \infty} \sum_{i=1}^{n-1} [P(A_{i+1}) - P(A_i)] \quad (\text{serie telescop}) \\ &= P(A_1) + \lim_{n \rightarrow \infty} [P(A_n) - P(A_1)] = \lim_{n \rightarrow \infty} P(A_n) \end{aligned}$$

Para las familias decrecientes, tomar complementos y usar este resultado (ejercicio).



Lema 1.2, continuación

7) σ -subaditividad: La probabilidad de la unión es menor o igual que la suma de las probabilidades:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i).$$

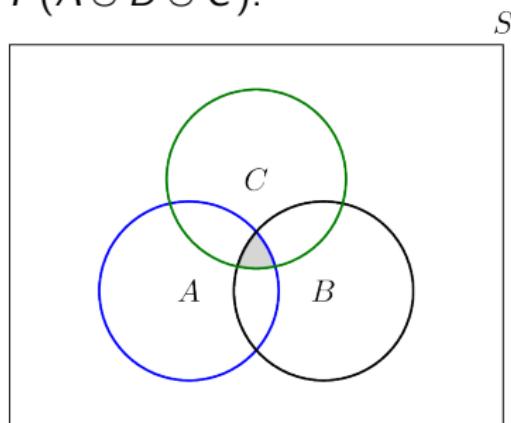
Demostración.

Cualquier union $\bigcup_{i=1}^{\infty} A_i$ puede escribirse como una unión de conjuntos disjuntos: eliminando de A_i la parte de A_i que ya está contenida en un A_j “anterior”.

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} A_i\right) &= P\left(\bigcup_{i=1}^{\infty}\left(A_i - \bigcup_{j=1}^{i-1} A_j\right)\right) \\ &\stackrel{\text{(disj)}}{=} \sum_{i=1}^{\infty} P\left(A_i - \bigcup_{j<i} A_j\right) \\ &\stackrel{\text{(monot)}}{\leq} \sum_{i \geq 1} P(A_i) \end{aligned}$$

Propiedades de la probabilidad: unión de 3 eventos

Si en lugar de dos eventos, tuviésemos tres, ¿cómo se podría calcular la probabilidad de la unión? Es decir, ¿cómo sería la fórmula de $P(A \cup B \cup C)$.



$$\begin{aligned}P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\&\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\&\quad + P(A \cap B \cap C)\end{aligned}$$

Lema 1.3 (Principio de inclusión-exclusión)

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2}) + \sum_{i_1 < i_2 < i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) \\ &\quad + \cdots + (-1)^{r+1} \sum_{i_1 < \cdots < i_r} P(A_{i_1} \cap \dots \cap A_{i_r}) + \dots \\ &= \sum_{k=1}^n \sum_{\substack{\mathcal{J} \subseteq \{1, \dots, n\}: \\ \#(\mathcal{J})=k}} (-1)^{k+1} P(\bigcap_{i \in \mathcal{J}} A_i) \end{aligned}$$

donde, por ejemplo, $\sum_{i_1 < i_2 < i_3}$ indica la sumatoria sobre todas las posibles combinaciones de los tres índices, i_1, i_2, i_3 cumpliendo las siguientes condiciones: $1 \leq i_1 < i_2 < i_3 \leq n$.

Demostración.

¡Ejercicio de la práctica!



Espacios Finitos o Numerables, $\Omega = \{\omega_i : i \in \mathbb{N}\}$

En este caso, todo evento puede ser pensado como una unión finita o numerable de sus elementos.

$$A \subset \Omega, \quad A = \bigcup_{\omega_i \in A} \{\omega_i\},$$

Luego, por aditividad o σ -aditividad de la probabilidad, (los conjuntos $\{\omega_i\}$ son disjuntos) tenemos

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\}) \quad (\text{serie absolutamente convergente})$$

Es decir,

Teorema 1.9

Toda probabilidad en un espacio muestral finito o numerable queda determinada por el valor de ésta en los eventos elementales.

Espacios Finitos o Numerables, $\Omega = \{\omega_i : i \in \mathbb{N}\}$

Teorema 1.10

Toda probabilidad en un espacio muestral finito o numerable queda determinada por el valor de ésta en los eventos elementales.

Demostración.

Sea $\{p_i : i \in \mathbb{N}\}$ una sucesión en $[0, 1]$ tal que $\sum_{i=1}^{\infty} p_i = 1$. A partir de ella definimos la función $P : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$ de modo que si $A = \bigcup_{\omega_i \in A} \{\omega_i\}$ entonces

$$P(A) = \sum_{i: \omega_i \in A} p_i,$$

o lo que es lo mismo, $P(\{\omega_i\}) = p_i$.



Espacios equiprobables. Espacios Muestrales finitos

Pensemos en el dado. El espacio muestral es $\{1, 2, 3, 4, 5, 6\}$ y si el dado está equilibrado, todos los valores tienen la misma chance de ocurrir. Tenemos entonces que $P(\{i\}) = c$ para $i = 1, \dots, 6$. $c = \frac{1}{6}$

Lema 1.4

En un espacio equiprobable $\Omega = \{\omega_1, \dots, \omega_N\}$, tenemos que

$$P(\{\omega_i\}) = \frac{1}{N} \text{ para todo } i = 1, \dots, N .$$

Demostración.

$$P(\{\omega_i\}) = c, \quad \text{para } i = 1, \dots, N .$$

$$1 = P(\Omega) = \sum_{i=1}^N P(\{\omega_i\}) = N c ,$$

de donde deducimos que c debe ser igual a $1/N$.

□

Espacios equiprobables. Espacios Muestrales finitos

Luego, si Ω es equiprobable y $A \subset \Omega$, $A = \bigcup_{\omega \in A} \{\omega\}$ entonces

$$P(A) = \frac{\#A}{N} = \frac{\text{cantidad de casos favorables}}{\text{total de casos}} ,$$

donde $\#A$ indica el cardinal de A , es decir, la cantidad de elementos de A .

Luego, en **espacios equiprobables** para calcular probabilidades hay que **saber contar**.

Ejemplo 1.4

Elegimos una permutación de los dígitos $\{0, 1, 2, \dots, 9\}$ al azar entre todas las disponibles.

- a) ¿Cuál es la probabilidad de que aparezca el número “1984” (como el nombre de la novela de Orwell) escrito en forma consecutiva?
Por ejemplo, la permutación 0198456723 lo tiene, y la sucesión 0197845623 no.
- b) ¿Cuál es la probabilidad de elegir una permutación que tenga el par 19, o bien el par 98 o bien el par 84?

Ejemplo 1.4 (solución) Permutaciones de dígitos

Elegimos una permutación de los dígitos $\{0, 1, 2, \dots, 9\}$ al azar entre todas las disponibles.

- a) ¿Cuál es la probabilidad de que aparezca el número “1984” (como el nombre de la novela de Orwell) escrito en forma consecutiva?

Por ejemplo, la permutación 0198456723 lo tiene, y la sucesión 0197845623 no. Definimos el siguiente espacio muestral $\Omega = \{\text{permutaciones de } \{0, 1, 2, \dots, 9\}\}$. El espacio es equiprobable. Por lo que calcular probabilidades se hace contando. $\#\Omega = 10!$ Definimos el evento

$$\begin{aligned} A_{1984} &= \{\text{permutaciones que contienen } 1984\} \\ &= \{\text{permutaciones de los 7 símbolos } \{0, \underline{1984}, 2, 3, 5, 6, 7\}\}. \end{aligned}$$

Luego $\#A_{1984} = 7!$, y tenemos

$$P(A_{1984}) = \frac{\#\text{casos favorables}}{\#\text{casos posibles}} = \frac{7!}{10!} = \frac{1}{10 \cdot 9 \cdot 8} \approx 0.00139$$

Ejemplo 1.4 (solución) Permutaciones de dígitos

b) ¿Cuál es la probabilidad de elegir una permutación que tenga el par 19, o bien el par 98 o bien el par 84?

Definamos, análogamente a A_{1984} los eventos A_{19} , A_{98} y A_{84} . Queremos calcular la $P(A_{19} \cup A_{98} \cup A_{84})$. Usamos el Principio de inclusión – exclusión,

$$\begin{aligned} P(A_{19} \cup A_{98} \cup A_{84}) &= P(A_{19}) + P(A_{98}) + P(A_{84}) \\ &\quad - [P(A_{19} \cap A_{98}) + P(A_{98} \cap A_{84}) + P(A_{19} \cap A_{84})] \\ &\quad + P(A_{19} \cap A_{98} \cap A_{84}) \end{aligned}$$

$$\begin{aligned} \#A_{19} &= \#\{\text{permutaciones de los 9 símbolos } \{0, \underline{19}, 2, 3, 4, 5, 6, 7, 8\}\} = 9! = \\ \#A_{98} &= \#A_{84}. \end{aligned}$$

Observemos que $A_{19} \cap A_{98} = A_{198}$ y

$$\#A_{198} = \#\{\text{permutaciones de los 8 símbolos } \{0, \underline{198}, 2, 3, 4, 5, 6, 7\}\} = 8!$$

También $A_{98} \cap A_{84} = A_{984}$ y $\#A_{984} = 8!$ Pero

$$\#(A_{19} \cap A_{84}) = \#\{\text{permutaciones de los 8 símbolos } \{\underline{19}, \underline{84}, 0, 2, 3, 5, 6, 7\}\} = 8!$$

El último término, $A_{19} \cap A_{98} \cap A_{84} = A_{1984}$, luego $P(A_{1984}) = \frac{7!}{10!}$ por el primer ítem.

Finalmente,

$$\begin{aligned} P(A_{19} \cup A_{98} \cup A_{84}) &= P(A_{19}) + P(A_{98}) + P(A_{84}) \\ &\quad - [P(A_{19} \cap A_{98}) + P(A_{98} \cap A_{84}) + P(A_{19} \cap A_{84})] \\ &\quad + P(A_{19} \cap A_{98} \cap A_{84}) \\ &= 3 \frac{9!}{10!} - 3 \frac{8!}{10!} + \frac{7!}{10!} \\ &= \frac{7!}{10!} [3 \cdot 9 \cdot 8 - 3 \cdot 8 + 1] = \frac{1}{10 \cdot 9 \cdot 8} 193 = 0.268 \end{aligned}$$

es decir, más de un cuarto de las veces.

Un ejercicio

Ejemplo 1.5

Una pequeña comunidad consta de 20 familias, 4 de las cuales tiene un único hijo, ocho familias tienen dos hijos, cinco familias tienen tres hijos, dos tienen cuatro hijos y una tiene 5 hijos.

- Si se selecciona una familia al azar, ¿cuál es la probabilidad de que tenga i hijos, para $i = 1, \dots, 5$?
- Si se selecciona un niño al azar, ¿cuál es la probabilidad de que pertenezca a una familia con i hijos, para $i = 1, 2, \dots, 5$?

Más sobre σ -álgebras

¿Cómo se puede construir una σ -álgebra? Sea \mathcal{G} una colección de subconjuntos de Ω . Definimos $\sigma(\mathcal{G})$ por

$$\sigma(\mathcal{G}) = \bigcap_{\substack{\mathcal{F} : \mathcal{F} \text{ es } \sigma\text{-álgebra} \\ \mathcal{G} \subset \mathcal{F}}} \mathcal{F}$$

Lema 1.5

$\sigma(\mathcal{G})$ es una σ -álgebra.

Demostración.

$\sigma(\mathcal{G})$ está bien definida pues la colección de todas las σ -álgebras \mathcal{F} en Ω que satisfacen $\mathcal{F} \supset \mathcal{G}$ es no vacía ya que $\mathcal{P}(\Omega)$ está incluída en ella. Como cada \mathcal{F} es una σ -álgebra:

- $\emptyset \in \mathcal{F}, \forall \mathcal{F}$, entonces $\emptyset \in \sigma(\mathcal{G})$
- Si $A \in \sigma(\mathcal{G})$ entonces $A \in \mathcal{F}, \forall \mathcal{F}$ entonces $A^c \in \mathcal{F}, \forall \mathcal{F}$ y por lo tanto $A^c \in \sigma(\mathcal{G})$
- Análoga prueba para la unión numerable.

σ -álgebra generada

Corolario 1.11

$\sigma(\mathcal{G})$ es la menor σ -álgebra que contiene a \mathcal{G} .

Definicion 1.12

$\sigma(\mathcal{G})$ se denomina la σ -álgebra generada por \mathcal{G} , y a \mathcal{G} lo llamamos un generador de $\sigma(\mathcal{G})$.

Miremos dos ejemplos de esta construcción.

Ejemplo 1.6

Sea Ω numerable y sea $\mathcal{G} = \{\{\omega\} : \omega \in \Omega\}$ la colección formada por los eventos elementales de Ω . Entonces, $\sigma(\mathcal{G}) = \mathcal{P}(\Omega)$. Esto se debe a que cada $A \in \mathcal{P}(\Omega)$ es numerable, entonces, a partir de la propiedad (3) de las σ -álgebras, se deduce que $A = \bigcup_{\omega \in A} \{\omega\} \in \sigma(\mathcal{G})$.

Ejemplo 1.7

(La σ -álgebra de Borel). Sea $\Omega = \mathbb{R}^n$ y

$$\mathcal{G} = \left\{ \prod_{i=1}^n [a_i, b_i] : a_i < b_i, a_i, b_i \in \mathbb{Q} \right\}$$

la colección que contiene todas las cajas rectangulares en \mathbb{R}^n con vértices racionales y caras paralelas a los ejes. La σ -álgebra $\mathcal{B}^n := \sigma(\mathcal{G})$ se denomina la σ -álgebra de Borel en \mathbb{R}^n , y a los conjuntos $A \in \mathcal{B}^n$ se los llama conjuntos de Borel o boreelianos. En el caso de $n = 1$, suele escribirse \mathcal{B} en vez de \mathcal{B}^1 . La σ -álgebra de Borel es mucho más grande que lo que uno imagina a simple vista. En realidad, tenemos

- a) Todo subconjunto abierto $A \subset \mathbb{R}^n$ es boreiano.
- b) Todo subconjunto cerrado $A \subset \mathbb{R}^n$ es boreiano.

Boreelianos de \mathbb{R}^n

Todo subconjunto abierto $A \subset \mathbb{R}^n$ es boreliano.

Demostración.

Para ver esto, basta observar que todo $\omega \in A$ tiene una vecindad $Q \in \mathcal{G}$ con $Q \subset A$, de modo que $A = \bigcup_{Q \in \mathcal{G}, Q \subset A} Q$ es una unión numerable de conjuntos. Luego, la afirmación se deduce de la propiedad de que las σ -álgebras son cerradas bajo uniones numerables. □

Todo subconjunto cerrado $A \subset \mathbb{R}^n$ es boreliano.

Es consecuencia de la propiedad de que las σ -álgebras son cerradas bajo complementos

Boreelianos de \mathbb{R}^n

No es posible describir a \mathcal{B}^n de manera constructiva. En Análisis Real se estudiará con detalle a esta σ -álgebra. Basta con saber que es una colección tan grande de conjuntos como para contener a todos los conjuntos relevantes en la práctica, pero que es menor a $\mathcal{P}(\mathbb{R}^n)$. De hecho, en Análisis Real se prueba la existencia de un conjunto no boreliano. Ver, por ejemplo

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons, para más detalles.

Observación 2

Para $\emptyset \neq \Omega \subset \mathbb{R}^n$, la colección $\mathcal{B}_\Omega^n = \{A \cap \Omega : A \in \mathcal{B}^n\}$ es una σ -álgebra en Ω ; se denomina la σ -álgebra de Borel en Ω .

Boreelianos de \mathbb{R}

La σ -álgebra \mathcal{B} en \mathbb{R} también admite a la colección de semi rectas a izquierda cerradas, como generador

$$\mathcal{G}' = \{(-\infty, c] : c \in \mathbb{R}\}$$

Esto es consecuencia de que al ser subconjuntos cerrados, por (b) sabemos que $\mathcal{G}' \subset \mathcal{B}$ y por lo tanto, por la minimalidad de $\sigma(\mathcal{G}')$, $\sigma(\mathcal{G}') \subset \mathcal{B}$. Recíprocamente, $\sigma(\mathcal{G}')$ contiene todos los intervalos acotados abiertos a la izquierda y cerrados a la derecha $(a, b] = (-\infty, b] - (-\infty, a]$, y por lo tanto también a los intervalos compactos $[a, b] = \bigcap_{n \geq 1} (a - \frac{1}{n}, b]$, y por lo tanto también a la σ -álgebra \mathcal{B} generada por ellos.

\mathcal{B} también puede ser generado por las semi rectas a izquierda abiertas. Y también por las semi rectas a derecha abiertas o cerradas.

σ -álgebras en esta materia

En la mayoría de los casos, en lo que concierne a esta materia, la elección de \mathcal{F} es canónica.

- caso discreto: si Ω es a lo sumo numerable, podemos tomar $\mathcal{F} = \mathcal{P}(\Omega)$.
- caso real: si $\Omega \subset \mathbb{R}^n$, la elección natural es $\mathcal{F} = \mathcal{B}_\Omega^n$.

¿Por qué complicarse con otras σ -álgebras?

¿Por qué tendrá sentido discutir otras σ -álgebras que no sean $\mathcal{P}(\Omega)$ para definir una probabilidad? Observemos que la probabilidad P tiene por dominio una colección de subconjuntos de Ω , y no a Ω . ¿Por qué estaríamos interesados en un caso en el que dicha colección de conjuntos no fuera $\mathcal{P}(\Omega)$?

Ejemplo 1.8 (Básquet)

Consideremos el experimento aleatorio en el que le pedimos a un jugador que tire 4 veces a un aro de básquet. En una ciudad, se les pide a todos los jugadores federados de 16 años que realicen el experimento. No tenemos acceso a la base de datos obtenida, pero en el informe publicado figura la siguiente tabla de resultados.

Cantidad de tiros embocados	0	1	2	3	4
Proporción de jugadores	0.25	0.26	0.15	0.18	0.16

¿Qué espacios de probabilidad podemos asociar al experimento?

Ejemplo 1.8 Básquet (cont.)

Una posibilidad es tomar $\Omega_1 = \{0, 1, 2, 3, 4\}$. Llámemos

$$p_0 = 0.25 \quad p_1 = 0.26 \quad p_2 = 0.15 \quad p_3 = 0.18 \quad p_4 = 0.16$$

Como Ω_1 es finito y las probabilidades de los eventos elementales están dadas, podemos definir a

$$P_1 : \mathcal{P}(\Omega_1) \rightarrow [0, 1]$$

usando el Teorema 1.10, por

$$P_1(A) = \sum_{i \in A} p_i$$

Esto nos daría un espacio $(\Omega_1, \mathcal{P}(\Omega_1), P_1)$.

Ejemplo 1.8 Básquet (cont.)

Otra posibilidad sería pensar en el espacio

$$\Omega_2 = \{(x_1, x_2, x_3, x_4) : x_i \in \{0, 1\}\}$$

donde

$$x_i = \begin{cases} 1 & \text{si el jugador acertó el tiro } i\text{-ésimo} \\ 0 & \text{si no lo hizo} \end{cases}$$

Ω_2 es finito, tiene $2^4 = 16$ elementos. Observemos que si quisiéramos definir a

$$P_2 : \mathcal{P}(\Omega_2) \rightarrow [0, 1]$$

no tenemos información suficiente para asignarle probabilidad a algunos subconjuntos, por ejemplo a $\{(0, 0, 1, 0)\}$ aunque sí la tenemos para asignársela al subconjunto $\{(0, 0, 0, 0)\}$, que es p_0 .

Ejemplo 1.8 Básquet (cont.)

Sin embargo, podemos trabajar con Ω_2 y definir $P_2 : \mathcal{F} \rightarrow [0, 1]$ donde \mathcal{F} está dada por

$$\mathcal{F} = \left\{ \Omega, \emptyset, A_0, A_1, A_2, A_3, A_4, \bigcup_{h \in H} A_h \text{ con } H \subseteq \{0, \dots, 4\}, 2 \leq \#H \leq 4, \right\}$$

donde

$$A_0 = \{(0, 0, 0, 0)\} \quad (\text{exactamente cero aciertos})$$

$$A_1 = \{(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)\} \quad (\text{exactamente un acierto})$$

$$A_2 = \{(1, 1, 0, 0), (1, 0, 1, 0), (1, 0, 0, 1), (0, 1, 1, 0), (0, 1, 0, 1), (0, 0, 1, 1)\} \quad (\text{exactamente dos aciertos})$$

$$A_3 = \{(0, 1, 1, 1), (1, 0, 1, 1), (1, 1, 0, 1), (1, 1, 1, 0)\} \quad (\text{exactamente tres aciertos})$$

$$A_4 = \{(1, 1, 1, 1)\} \quad (\text{exactamente cuatro aciertos})$$

Es fácil comprobar que \mathcal{F} es una σ -álgebra (la σ -álgebra generada por los $\{A_i\}_{0 \leq i \leq 4}$) y que sabemos cómo definir la probabilidad en cada uno de los elementos de \mathcal{F} . Esto nos da otro espacio de probabilidad para este experimento: $(\Omega_2, \mathcal{F}, P_2)$.

Ejemplo 1.8 Básquet (cont.)

¿Cuál de los dos espacios de probabilidad es mejor para describir el experimento aleatorio? Ambos funcionan perfectamente bien, por ahora. Si obtuviéramos un poco más de información sobre el experimento, digamos que pudiéramos acceder a otro cuadro con un resumen parcial de los resultados, con la siguiente información

Embocaron los primeros tres tiros	Sí	No
Proporción de jugadores	0.21	0.79

Observemos que no tenemos cómo incorporar esta nueva información al espacio $(\Omega_1, \mathcal{P}(\Omega_1), P_1)$. Sin embargo, sí tenemos forma de incluirla en Ω_2 . Lo hacemos agrandando la σ -álgebra \mathcal{F} a una \mathcal{F}_2 que incluirá al evento

$$B = \{\text{embocaron los tres primeros tiros}\} = \{(1, 1, 1, 0), (1, 1, 1, 1)\}.$$

Luego, $\mathcal{F}_2 = \sigma(A_0, A_1, \dots, A_4, B)$. Como $\{(1, 1, 1, 0)\} \in \mathcal{F}_1$, le podemos asignar probabilidad ya que $\{(1, 1, 1, 0)\} = B - A_4$ y $A_4 \subset B$ resulta $P(\{(1, 1, 1, 0)\}) = P(B) - P(A_4) = 0.21 - 0.16 = 0.05$

Ejemplo 1.8 Básquet (cont.)

Entonces, con la nueva información podemos definir a P_2 en \mathcal{F}_2 y tenemos un espacio de probabilidad para el experimento, que modela toda la información disponible: $(\Omega_2, \mathcal{F}_2, P_2)$. \mathcal{F}_2 es una σ -álgebra más rica que \mathcal{F} ya que contiene más elementos (más subconjuntos de Ω).

La σ -álgebra codifica la información disponible

En probabilidad, las σ -álgebras se usan para codificar o representar la información disponible.

En ese sentido, si para el ejemplo del básquet dispusiéramos de la **base de datos completa**, tendríamos la posibilidad de asignarle probabilidad a **todos los subconjuntos de Ω_2** , ya que agrupando las respuestas de los distintos jugadores podríamos calcular, por ejemplo, la probabilidad de obtener $\{(1, 0, 0, 0)\}$ y la de $\{(0, 0, 0, 1)\}$.

Al compararlas podríamos responder a la pregunta de si **es más probable acertar en el primer tiro o en el cuarto (si hay un “efecto entrada en calor” o no lo hay)**.

Más información sobre el experimento aleatorio se traduce en la posibilidad de definir la probabilidad en una **σ -álgebra más detallada, más rica, con más eventos**, que permite discernir entre más resultados y responder más preguntas. Es en ese sentido es que debe leerse la **primera frase** de esta página.

Un último comentario. Cuando el espacio Ω es \mathbb{R} o \mathbb{R}^n , vimos que tomamos la σ -álgebra de los borelianos (de \mathbb{R} o \mathbb{R}^n , respectivamente). Mencionamos que es imposible definir una probabilidad en $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$. Para aquel que no pueda esperar la prueba que verán en Análisis Real, pueden consultar (con un nivel matemático más elevado que el tratado en este curso) el Teorema 1.5 (una versión probabilística de la construcción del Teorema de Vitali), en Georgii, Hans-Otto. *Stochastics, Introduction to Probability and Statistics*, Second edition, 2013. Sin embargo, les recomiendo postergar esta discusión para la materia correlativa.

2. Probabilidad condicional

La probabilidad asignada a un evento depende de lo que se sabe acerca de la situación experimental en el momento de calcularla.

Posteriormente a dicho cálculo, puede aparecer información adicional que permite reasignar mejor la probabilidad.

Si A es un evento, $P(A)$ es la probabilidad original del evento. ¿Cómo introducimos en el cálculo de la probabilidad de A la información de “ha ocurrido el evento B ”?

Usaremos la notación $P(A|B)$ para indicar la probabilidad de A sabiendo (o condicional a) que B ocurrió.

Ejemplo: Basado en la encuesta de **Amnistía Internacional** sobre la **despenalización del aborto** en Argentina, marzo 2018 que puede encontrarse en este link

https://amnistia.org.ar/wp-content/uploads/delightful-downloads/2018/03/Informe_aborto_130318-.pdf?utm_source=Prensa&utm_campaign=77d240c88c-EMAIL_CAMPAIGN_2018_03_16&utm_medium=email&utm_term=0_a60e315cac-77d240c88c-

Los números para este ejemplo fueron redondeados.

Colors - Informe_aborto_130318-p... x +

https://amnistia.org.ar/wp-content/uploads/delightful-downloads/2018/03/informe_aborto_130318-...

Automatic Zoom

1 of 40

SITUACIÓN DE LA OPINIÓN PÚBLICA CON RESPECTO AL ABORTO

INFORME DE RESULTADOS

Marzo 2018

AMNISTÍA INTERNACIONAL

CEDES

! QUIDDITY

LUIS COSTA
Luis@quiddity.info

1

Type here to search

Windows Start button

e-mail

File Explorer

Firefox

Microsoft Edge

OneDrive

OneNote

PowerPoint

Photoshop

PDF

Recycle Bin

Search

Task View

Teams

Word

YouTube

File

Print

Save

Stop

Forward

Back

Home

Search

Help

ENGLISH

INTL

5:32 PM

06-Apr-20

Ejemplo 2.1

Una consultora de opinión realiza una encuesta sobre 1600 adultos elegidos al azar en marzo 2018, en Argentina. Les pregunta su opinión sobre la despenalización del aborto, y obtiene los siguientes resultados

	edad		total
	≤ 45 años	> 45 años	
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

(basado en una encuesta de *Amnistía Internacional*)

Ejemplo despenalización del aborto

	edad		total
	≤ 45 años	> 45 años	
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

Se elige una persona al azar entre las encuestadas.

Queremos hallar la $P(A)$ siendo

$A = \{ \text{está de acuerdo con la despenalizac del} \}$
aborto

	edad		total
	≤ 45 años	> 45 años	
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

Ejemplo despenalización del aborto

Se elige una persona al azar entre las encuestadas.

Queremos hallar la $P(A)$ siendo

$A = \{ \text{está de acuerdo con la despenalizac del} \}$
aborto

$$P(A) = \frac{\# \text{ casos fav}}{\# \text{ casos posibles}} = \frac{930}{1600} = 0,58125$$

espacio
equip.

	edad		
	≤ 45 años	> 45 años	total
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

Ejemplo despenalización del aborto

Se elige una persona al azar entre las encuestadas.

Queremos hallar la $P(A)$ siendo

$A = \{ \text{está de acuerdo con la despenalizac del} \}$
aborto

$$P(A) = \frac{\# \text{casos fav}}{\# \text{casos posibles}} = \frac{930}{1600} = 0,58125$$

espacio

equip. Pero qué pasa ahora, si sabemos que la persona elegida es joven

	edad		
	≤ 45 años	> 45 años	total
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

Ejemplo despenalización del aborto

Se elige una persona al azar entre las encuestadas.

Queremos hallar la $P(A)$ siendo

$A = \{ \text{está de acuerdo con la despenalizac del} \}$
aborto

$$P(A) = \frac{\# \text{casos fav}}{\# \text{casos posibles}} = \frac{930}{1600} = 0,58125$$

espacio

equip. Pero qué pasa ahora, si sabemos que la persona elegida es joven. Sea

$$B = \{ \text{edad menor o igual a } 45 \text{ años} \}$$

Ahora nos importa $P(A | B)$. ¿será igual?

Ejemplo despenalización del aborto

	edad		
	≤ 45 años	> 45 años	total
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

¿cómo la calcularíamos?

los que están deacuerdo
y son jóvenes.

$$P(A|B) = \frac{\# \text{ casos fav.}}{\# \text{ casos posibles}} =$$

restringimos el universo a los jóvenes.

Ejemplo despenalización del aborto

	edad		
	≤ 45 años	> 45 años	total
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

¿cómo la calcularíamos?

los que están deacuerdo
y son jóvenes.

$$P(A|B) = \frac{\# \text{ casos fav.}}{\# \text{ casos posibles}} = \frac{530}{860} = 0,6163$$

(1)

restringimos el universo a los jóvenes.

Ejemplo despenalización del aborto

	edad		
	≤ 45 años	> 45 años	total
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

¿cómo la calcularíamos?

10.

los que están de acuerdo
y son jóvenes.

$$P(A|B) = \frac{\# \text{ casos fav.}}{\# \text{ casos posibles}} = \frac{530}{860} = 0,6163 \quad (1)$$

restringimos el universo a los jóvenes.

$$P(A|B) = \frac{\frac{530}{1600}}{\frac{860}{1600}} = \frac{P(A \cap B)}{P(B)} \quad (2)$$

Vemos que el dato de la
edad modifica el cálculo de la probab.

	edad		total
	≤ 45 años	> 45 años	
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

Ejemplo despenalización del aborto

Resumiendo,

$$P(A) = 0,58125$$

$$P(A|B) = 0,6163$$

Entre los jóvenes hay **mayor** probabilidad de estar de acuerdo con la despenalización del aborto

Definición de probabilidad condicional

Definición: Si B es un evento con $P(B) > 0$, entonces

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Es la probabilidad de A condicional a que B ocurrió.

También se la llama *probabilidad de A sabiendo que B ocurrió*, o también *probabilidad de A dado que B ocurrió*.

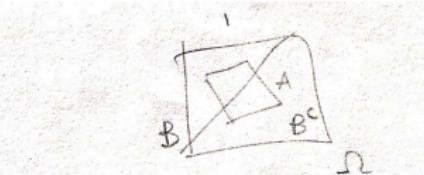
	edad		total
	≤ 45 años	> 45 años	
de acuerdo	530	400	930
en desacuerdo	300	300	600
no sabe	30	40	70
total	860	740	1600

Ejemplo despenalización del aborto

$$P(A|B) \text{ no es } P(A \cap B)$$

En el ejemplo:

$$P(A \cap B) = \frac{530}{1600} = 0,33$$



elijo una persona y me pregunto si es joven y si está de acuerdo

$$P(A|B) = 0,62$$

elijo una persona joven y me pregunto si está de acuerdo

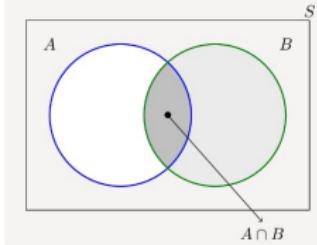
$P(\cdot|B)$ es una probabilidad

Entonces $P(\cdot|B)$ es una nueva probabilidad definida sobre los eventos del espacio muestral Ω , que vale 1 sobre el evento B y cero en todo evento A disjunto de B .

Lema 2.1

$P(\cdot|B)$ es una función de probabilidad en (Ω, \mathcal{F}) .

Al revés NO, $P(A|\cdot)$ no es una probabilidad.



Lema 2.2

$P(\cdot|B)$ es una función de probabilidad en (Ω, \mathcal{F}) .

Demostración.

1) $P[\Omega|B] = \frac{P[\Omega \cap B]}{P[B]} = \frac{P[B]}{P[B]} = 1$

2) Sean $A_i \in \mathcal{F} \forall i \in \mathbb{N}$, $A_i \cap A_j = \emptyset \quad \forall i \neq j$,

$$\begin{aligned} P\left[\bigcup_{i=1}^{\infty} A_i | B\right] &= \frac{P[(\bigcup_{i=1}^{\infty} A_i) \cap B]}{P[B]} = \frac{P[\bigcup_{i=1}^{\infty} (A_i \cap B)]}{P[B]} \text{ (distributiva)} \\ &= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P[B]} \quad (\sigma\text{-adit para } P, A_i \cap B \text{ son disj}) \\ &= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P[B]} = \sum_{i=1}^{\infty} P(A_i | B) \quad (\text{def de prob cond}) \end{aligned}$$



Ejemplo 2.2

Una familia tiene dos hijos. Sabiendo que al menos uno es varón, ¿cuál es la probabilidad de que ambos sean varones?

Armemos el espacio muestral

$$\Omega = \{(V, V), (V, M), (M, V), (M, M)\}$$

donde la primer coordenada representa el género del vástago mayor y el segundo el del menor, y "V" indica que se trata de un varón, mientras que "M" indica que no se trata de un varón.

Asumimos un **espacio equiprobable**, es decir

$$P((V, V)) = P((V, M)) = P((M, V)) = P((M, M)) = \frac{1}{4}$$

equiprobable

Nos interesan los eventos

$$A = \{(V, V)\}, \quad B = \{\text{al menos un varón}\} = \{(V, V), (V, M), (M, V)\}$$

Ejemplo 2.2 cont.

Una familia tiene dos hijos. Sabiendo que al menos uno es varón, ¿cuál es la probabilidad de que ambos sean varones?

$$\begin{aligned} P(A|B) &= P((V, V)|\{(V, V), (V, M), (M, V)\}) = \frac{P(A \cap B)}{P(B)} \\ &= \frac{P((V, V))}{P(B)} = \frac{1/4}{3/4} = \boxed{\frac{1}{3}} \end{aligned}$$

Ejercicio. Una familia tiene dos hijos. Calcule la probabilidad de que ambos sean varones, sabiendo que el mayor es varón. Esta respuesta no será contraintuitiva

Probabilidad de $A \cap B$

La definición de probabilidad condicional vincula la probabilidad condicional $P(A|B)$ con la probabilidad de una intersección $P(A \cap B)$. A veces la $P(A|B)$ es un dato disponible y podemos usarla para calcular la $P(A \cap B)$. De la definición tenemos que

$$P(A \cap B) = P(B)P(A|B) \text{ si } P(B) > 0 \quad (1)$$

$$= P(A)P(B|A) \text{ si } P(A) > 0 \quad (2)$$

Ejemplo 2.3

Tenemos una urna con 3 bolas rojas, 2 bolas azules y 2 bolas verdes. Se sacan 2 bolas sin reposición. Calcular la probabilidad de que ambas sean rojas.

Ejemplo 2.3 cont.

Tenemos una urna con 3 bolas rojas, 2 bolas azules y 2 bolas verdes. Se sacan 2 bolas **sin** reposición. Calcular la probabilidad de que ambas sean rojas.

Numeramos las bolas, 1,2,3,4,5,6,7

(las hacemos distinguibles entre sí)

$$\Omega = \{(x_1, x_2) : 1 \leq x_1, x_2 \leq 7, x_1 \neq x_2\}$$

x_i representa la i-ésima bola extraída

Sean $R_1 = \{1^{\text{er}} \text{ bola roja}\}$

$R_2 = \{2^{\text{a}} \text{ bola roja}\}$

Queremos $P(R_1 \cap R_2) = P(R_1) \cdot P(R_2 | R_1)$

$$P(R_1) = \frac{\#\text{casos favorables}}{\#\text{casos posibles}}$$



Ejemplo 2.3 cont.

Tenemos una urna con 3 bolas rojas, 2 bolas azules y 2 bolas verdes. Se sacan 2 bolas **sin** reposición. Calcular la probabilidad de que ambas sean rojas.

Numeramos las bolas, 1,2,3,4,5,6,7

(las hacemos distinguibles entre sí)

$$\Omega = \{(x_1, x_2) : 1 \leq x_1, x_2 \leq 7, x_1 \neq x_2\}$$

x_i representa la i-ésima bola extraída

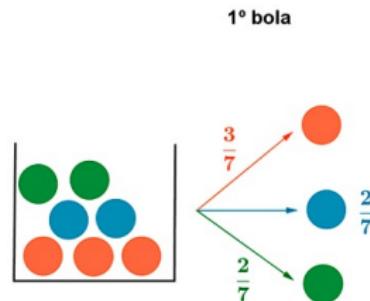
Sean $R_1 = \{1^{\text{er}} \text{ bola roja}\}$

$R_2 = \{2^{\text{a}} \text{ bola roja}\}$

$$\text{Queremos } P(R_1 \cap R_2) = P(R_1) \cdot P(R_2|R_1)$$

$$P(R_1) = \frac{\#\text{casos favorables}}{\#\text{casos posibles}} = \frac{3}{7}$$

$$P(R_2|R_1)$$



Ejemplo 2.3 cont.

Tenemos una urna con 3 bolas rojas, 2 bolas azules y 2 bolas verdes. Se sacan 2 bolas sin reposición. Calcular la probabilidad de que ambas sean rojas.

Numeramos las bolas, 1,2,3,4,5,6,7

(las hacemos distinguibles entre sí)

$$\Omega = \{(x_1, x_2) : 1 \leq x_1, x_2 \leq 7, x_1 \neq x_2\}$$

x_i representa la i-ésima bola extraída

Sean $R_1 = \{1^{\text{er}} \text{ bola roja}\}$

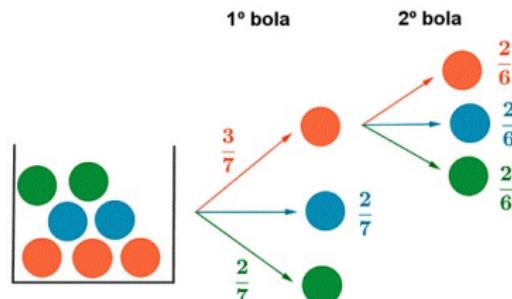
$R_2 = \{2^{\text{a}} \text{ bola roja}\}$

Queremos $P(R_1 \cap R_2) = P(R_1) \cdot P(R_2|R_1)$

$$P(R_1) = \frac{\#\text{casos favorables}}{\#\text{casos posibles}} = \frac{3}{7}$$

$$P(R_2|R_1) = \frac{\#\text{casos favorables}}{\#\text{casos posibles}} = \frac{2}{6} \quad (\text{pensar en una urna nueva con 6 bolas})$$

Finalmente, $P(R_1 \cap R_2) = \frac{3}{7} \cdot \frac{2}{6} = \frac{1}{7}$



Ejemplo 2.3 cont.

Tenemos una urna con 3 bolas rojas, 2 bolas azules y 2 bolas verdes. Se sacan 3 bolas **sin** reposición. Calcular la probabilidad de observar **roja verde roja** (en ese orden).

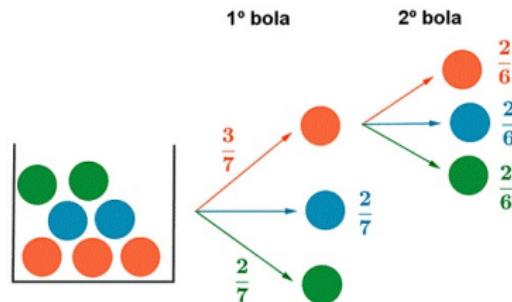
Ahora sacamos 3 bolas, sean

$$R_1 = \{1^{\text{er}} \text{ bola roja}\}$$

$$V_2 = \{2^{\text{a}} \text{ bola verde}\}$$

$$R_3 = \{3^{\text{er}} \text{ bola roja}\}$$

$$\begin{aligned} P(R_1 \cap V_2 \cap R_3) &= P((R_1 \cap V_2) \cap R_3) \\ &= P(R_1 \cap V_2) \cdot P(R_3 | R_1 \cap V_2) \\ &= P(R_1) \cdot P(V_2 | R_1) \cdot P(R_3 | V_2 \cap R_1) \\ &= \frac{3}{7} \cdot \frac{2}{6} \cdot \frac{2}{5} = \frac{2}{35} \end{aligned}$$



Este ejemplo contiene una lección importante. Fuimos capaces de calcular la probabilidad de una intersección sin calcular primero las probabilidades de cada evento de Ω , simplemente usando la definición de probabilidad condicional. Esto es una técnica habitual en el área: muchas veces no es necesario preocuparse por el espacio muestral o calcular la función de probabilidad en todos los resultados posibles.

Podemos extender la regla multiplicativa a varios eventos

Lema 2.3

Sean $A_1, A_2, \dots, A_n \in \mathcal{F}$ tales que $P(A_1 \cap \dots \cap A_n) > 0$, entonces

$$P(A_1 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\dots P(A_n|A_1 \cap \dots \cap A_{n-1})$$

Demostración.

Por inducción en n . Caso $n = 2$ ya probado.

Paso inductivo: $\mathcal{P}(n-1) \Rightarrow \mathcal{P}(n)$: Observemos que si $P(A_1 \cap \dots \cap A_n) > 0$, entonces por monotonía, $P(A_1 \cap \dots \cap A_{n-1}) > 0$. Entonces

$$P(A_1 \cap \dots \cap A_n) = P((A_1 \cap A_2) \cap A_3 \cap \dots \cap A_n)$$

$$\stackrel{\text{H.I.}}{=} P(A_1 \cap A_2)P(A_3|A_1 \cap A_2)\dots P\left(A_n|\bigcap_{i=1}^{n-1} A_i\right)$$

$$\stackrel{\text{caso } n=2}{=} P(A_1)P(A_1|A_2)P(A_3|A_1 \cap A_2)\dots P\left(A_n|\bigcap_{i=1}^{n-1} A_i\right).$$



Ejemplo 2.4 (Seguros)

Una compañía aseguradora divide a sus clientes en tres clases: alto riesgo, mediano riesgo, y bajo riesgo. Su cartera de clientes está compuesta por un 20 % de clientes de alto riesgo, 30 % de mediano riesgo y 50 % de bajo riesgo. La probabilidad de que un cliente dado tenga un accidente en el corriente año es 0.25 para los de alto riesgo, 0.16 para los de mediano riesgo, y 0.10 para los de bajo riesgo.

- Encuentre la probabilidad de que un cliente elegido al azar, tenga un accidente en el corriente año.
- Halle la probabilidad de que un cliente sea de alto riesgo, dado que se sabe que ha tenido un accidente durante el año corriente.

El espacio muestral $\Omega = \{\text{clientes de la compañía}\}$. Se definen los eventos:

$A_1 = \{\text{el cliente es de alto riesgo}\}$ $A_2 = \{\text{el cliente es de mediano riesgo}\}$

$A_3 = \{\text{el cliente es de bajo riesgo}\}$ $E = \{\text{el cliente sufre un accidente}\}$

$A_1 \cup A_2 \cup A_3 = \Omega$ y son disjuntos de a pares. Los datos son

$$P(E|A_1) = 0,25; \quad P(A_1) = 0,2$$

$$P(E|A_2) = 0,16; \quad P(A_2) = 0,3$$

$$P(E|A_3) = 0,10; \quad P(A_3) = 0,5$$

$$a) P(E) = P(E \cap \Omega) = P(E \cap (A_1 \cup A_2 \cup A_3)) \underset{\substack{= \\ (\text{distributiva})}}{=} P\left(\bigcup_{i=1}^3 (E \cap A_i)\right)$$

$$\underset{\substack{= \\ (\text{disj})}}{\sum_{i=1}^3} P(E \cap A_i) = \sum_{i=1}^3 P(A_i)P(E|A_i)$$

$$= 0,20 \times 0,25 + 0,3 \times 0,16 + 0,5 \times 0,10 = 0,148$$

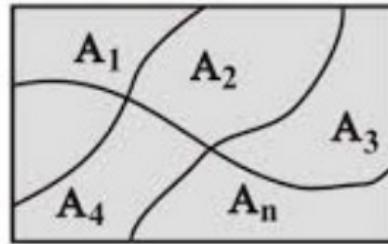
$$b) P(A_1|E) = \frac{P(A_1 \cap E)}{P(E)} = \frac{P(A_1)P(E|A_1)}{P(E)} = \frac{0,20 \times 0,25}{0,148} = 0,3378 \text{ (¡aumenta!)}$$

Ahora formalizamos lo hecho para el ejemplo de seguros.

Definición 1

Diremos que los eventos $(A_i)_{i \geq 1}$ forman una *partición del espacio muestral* Ω si

- i. Los eventos son disjuntos dos a dos: $A_i \cap A_j = \emptyset$, para $i \neq j$.
- ii. Los eventos cubren el espacio muestral: $\bigcup_{i \geq 1} A_i = \Omega$.

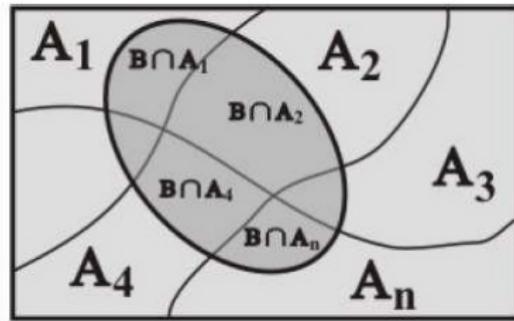


Teorema 2.1

(Probabilidad Total) Dada una partición $(A_i)_{i \geq 1}$ del espacio muestral Ω , con $A_i \in \mathcal{F}$ y $P(A_i) > 0$ para todo i , tenemos que

$$P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).$$

En realidad, basta con que los eventos A_i sean disjuntos y $\sum_i P(A_i) = 1$.



Teorema 2.2

(Regla de Bayes) Dada una partición $(A_i)_{i \geq 1}$ del espacio muestral Ω , tenemos que

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}, \text{ para } k \geq 1.$$

Este teorema tiene aplicaciones importantes: permite “dar vuelta los condicionales”, relacionando las $P(B|A_i)$ y $P(A_i|B)$. Los nombres que se les suelen dar a estas probabilidades en algunos contextos:

$P(A_i)$: probabilidades *a priori* de los elem de la partición

$P(A_i|B)$ probabilidades *a posteriori* de los elem de la partición,

una vez que se incorporó la información de que B ocurrió

Demostración.

Elemental usando la def de prob condic y el teo de la probabilidad total

Animación sobre Teorema de Bayes

www.3blue1brown.com

Alucinante video, en este link

<https://youtu.be/HZGCoVF3YvM>

*Perhaps the most important formula in probability
Bayes theorem, and making probability intuitive
(en inglés pero subtulado)*

Y su breve continuación

https://youtu.be/U_85TaXbeIo

The quick proof

Ejemplo 2.5

Cuando se realiza un análisis de laboratorio para diagnosticar una cierta enfermedad en un paciente se está frente a la posibilidad de cometer **dos tipos de errores** en el diagnóstico. Sean

$E = \{ \text{la persona examinada está enferma} \}$ y

$A = \{ \text{el resultado del análisis es positivo} \}$, es decir que el análisis concluye que la persona examinada contrajo la enfermedad.

Por supuesto, ambos eventos no tienen por qué coincidir (aunque eso sería lo deseable). Cuando no lo hacen, hay un error de diagnóstico: si el análisis da positivo pero el paciente está sano se dice que tenemos un falso positivo, si en cambio el análisis da negativo pero el paciente está enfermo se dice que tenemos un falso negativo. Para cada análisis se conocen la

sensibilidad del test $\rightarrow P(A|E)$

especificidad del test $\rightarrow P(A^c|E^c)$

		A (análisis +)	A^c (análisis -)
verdad	E (enferma)	✓	falso negativo
	E^c (sana)	falso positivo	✓

Continuación Ejemplo 2.5

Supongamos que una prueba de laboratorio en particular es tal que

$$P(A|E) = 0,95 \quad P(A^c|E^c) = 0,80$$

y que la probabilidad de que una persona padezca la enfermedad en esta población es 0,001 (prevalencia de la enfermedad). ¿Cuál es la probabilidad de que una persona cuyo análisis diagnóstico es positivo en realidad esté enferma? Interpretar la respuesta.

Continuación Ejemplo 2.5

Supongamos que una prueba de laboratorio en particular es tal que

$$P(A|E) = 0,95 \quad P(A^c|E^c) = 0,80$$

y que la probabilidad de que una persona padezca la enfermedad en esta población es 0,001 (prevalencia de la enfermedad). ¿Cuál es la probabilidad de que una persona cuyo análisis diagnóstico es positivo en realidad esté enferma? Interpretar la respuesta.

Observar que $\{E, E^c\}$ es una partición de Ω

$$\begin{aligned} P(E|A) &= \frac{P(A|E)P(E)}{P(A|E)P(E) + P(A|E^c)P(E^c)} \\ &= \frac{(0,95)(0,001)}{(0,95)(0,001) + (0,2)(0,999)} = 4,7 \times 10^{-3} \end{aligned}$$

¡La probabilidad de estar enfermo sabiendo que el test dio + es muy baja! Esto es porque la mayor parte de la población está sana. De hecho, tener un test positivo multiplica casi por 5 la probabilidad de estar enfermo.

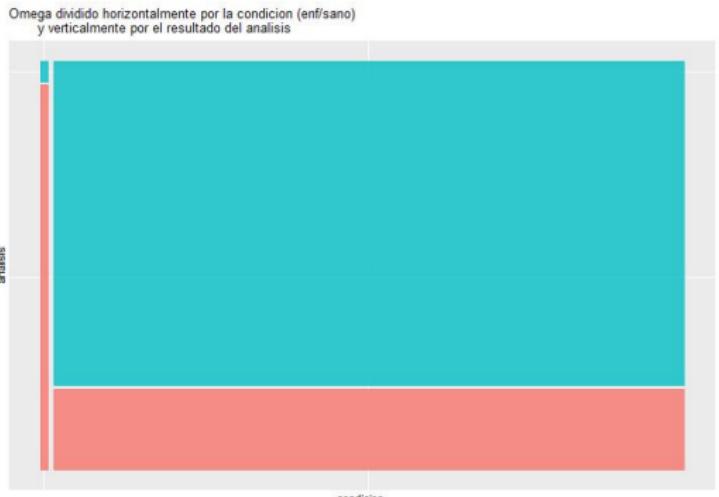
Ejemplo: Test diagnóstico (cont.) (gráficos con ggmosaic de R)

Graficamos la partición $\{E, E^c\}$ de Ω , los enfermos (a la izquierda) representan una pequeña proporción del total (pues $P(E)$ es pequeña).



Ejemplo: Test diagnóstico (cont.)

Graficamos la partición $\{E, E^c\}$ de Ω , los enfermos representan una pequeña proporción del total (pues $P(E)$ es pequeña). Le agregamos el evento **A (test positivo)**. Como la $P(A|E)$ es mayor que $P(A|E^c)$, la región **rosa** ocupa en el gráfico una proporción mayor de E que de E^c . Así y todo, una vez que sabemos que el test dio positivo (es decir, si sabemos que hemos elegido un ω en el espacio muestral que es rosa) tiene más chance de estar en la zona de las personas sanas que de las enfermas.



Ejemplo: Test diagnóstico (cont.)

Si en cambio el médico le hubiera mandado a hacer el análisis en cuestión **a un paciente quien se queja de síntomas consistentes con la enfermedad** (es decir, pertenece a la población de riesgo) y podemos asumir que $P(E) = 0,5$, entonces

$$P(E|A) = \frac{(0,95)(0,5)}{(0,95)(0,5) + (0,2)(0,5)} = 0,83$$

Es por eso que **no** es una buena política pública de diagnóstico de una enfermedad poco prevalente el hacer tests masivos sobre la población general.

Independencia

Tras haber definido la probabilidad condicional, ¿cuándo diría que dos eventos son independientes? Podríamos decir que dos eventos son independientes si conocer la ocurrencia de unos de ellos nada nos dice sobre la ocurrencia del otro. En otras palabras, podemos decir que dos eventos A y B son independientes cuando $P(A|B) = P(A)$. Recordando la definición de probabilidad condicional tenemos que:

Definición 2

Dos eventos A y $B \in \mathcal{F}$ se dicen *independientes* si

$$P(A \cap B) = P(A)P(B).$$

Ojo, no confundir eventos independientes con disjuntos. Recordemos que A y B son disjuntos si $A \cap B = \emptyset$.

Ejemplo 2.6

Elegimos una carta al azar de un mazo de 48 cartas españolas. Sean

$$A = \{\text{la carta es de espada}\}$$

$$B = \{\text{la carta es un as}\}$$

Entonces $P(A) = \frac{12}{48} = \frac{1}{4}$ $P(B) = \frac{4}{48} = \frac{1}{12}$.

El evento

$$A \cap B = \{\text{la carta es el as de espada}\}$$

satisface

$$P(A \cap B) = \frac{1}{48} = \frac{1}{4} \cdot \frac{1}{12} = P(A) \cdot P(B)$$

Como $P(A \cap B)$ resulta ser igual a $P(A) \cdot P(B)$, afirmamos que los eventos A y B son independientes.

En otras palabras, si sabemos que salió un as, *no sabemos nada* sobre si se trata del as de espadas o no.

Propiedades de independencia

Lema 2.4

Sean A y $B \in \mathcal{F}$.

- A y B son independientes, entonces A y B^c son independientes.
- A es independiente de sí mismo si y solo si $P(A) = 0$ ó $P(A) = 1$.

Si aplicamos el Lema anterior una vez más, obtenemos que A^c y B^c son independientes y lo mismo vale para A^c y B .

Demostración.

- El primer resultado se generaliza en un ejercicio de la práctica.

$$\begin{aligned} P(A \cap B^c) &= P(A - A \cap B) \underset{(A \cap B \subset A)}{=} P(A) - P(A \cap B) \\ &\underset{(A \text{ y } B \text{ indep})}{=} P(A) - P(A)P(B) = P(A)[1 - P(B)] = P(A)P(B^c). \end{aligned}$$

- Es un ejercicio de la práctica. □

Observación 1

A veces, la independencia entre eventos puede asumirse, es decir, forma parte de las hipótesis del problema.

- **Por ejemplo**, cuando tiramos dos veces consecutivas una moneda podemos asumir que el resultado del primer tiro no condiciona el del segundo (y al revés).

Sin embargo, a veces para decidir si dos eventos son o no independientes es necesario primero hacer la cuenta y luego concluir.

- **Por ejemplo**, cuando probamos que los eventos $\{\text{sacar un as}\}$ y $\{\text{sacar una espada}\}$ son independientes en el experimento de sacar una carta al azar de un mazo de cartas españolas.

Independencia de más de dos eventos

Generalizamos el concepto de independencia para cuando tenemos más de dos eventos. Como en el caso de dos eventos, la idea es que al condicionar, obtengamos las mismas probabilidades. Comenzamos pidiendo que $P(A_i|A_j) = P(A_i)$, de donde obtenemos que $P(A_i \cap A_j) = P(A_i)P(A_j)$. Luego pedimos que $P(A_i|A_j \cap A_k) = P(A_i)$, de donde deducimos que $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_k \cap A_j)$. Pero como ya sabemos que para intersecciones de a dos la probabilidad se factoriza, obtenemos que $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$. Continuando con este argumento, resulta natural que la definición de eventos independientes pida que la probabilidad de cualquier intersección de algunos de tales eventos, coincida con el producto de las probabilidades de los eventos intersecados.

Independencia de más de dos eventos

Definición 3

Una familia de eventos $(A_t)_{t \in T}$, $A_t \in \mathcal{F}$ se dice independiente si para todo $F \subset T$ finito, vale que

$$P\left(\bigcap_{t \in F} A_t\right) = \prod_{t \in F} P(A_t).$$

Independencia e independencia de a pares

Observe que independencia dos a dos de los eventos no garantiza independencia general, tal como lo muestra el siguiente ejemplo.

Ejemplo 2.7

Tire dos veces una moneda y considere los siguientes eventos:

- a) A : sale cara la primera vez, entonces $P(A) = 1/2$.
- b) B : sale cara la segunda vez, entonces $P(B) = 1/2$.
- c) C : sale cara exactamente una vez, entonces $P(C) = 1/2$.

Calcule las siguientes probabilidades para verificar que los eventos son independientes dos a dos.

$$P(A \cap B) = \dots$$

$$P(A \cap C) = \dots$$

$$P(B \cap C) = \dots$$

¿Cuánto vale $P(A \cap B \cap C) = \dots$? ¿Puede decir que los tres eventos sean independientes?

Despenalización, condiciones para independencia

Volvamos a la encuesta de despenalización del aborto.

Ejemplo 2.8

Si mantenemos los totales encuestados en cada grupo

	edad		total
	≤ 45 años	> 45 años	
de acuerdo		930	
en desacuerdo		600	
no sabe		70	
total	860	740	1600

¿cómo debe ser completada esta tabla para que los eventos $A = \{\text{está de acuerdo}\}$ y $B = \{45 \text{ años o menos}\}$ sean independientes y que $C = \{\text{no sabe}\}$ y B también lo sean? (redondear al entero más cercano)

3. Variables Aleatorias

Muchas veces no nos interesarán todos los aspectos de un experimento aleatorio, sino que queremos concentrar nuestra atención en alguna consecuencia numérica del mismo.

Por ejemplo, para un apostador puede resultar más interesante la ganancia o pérdida obtenida, más que registrar el resultado del juego realizado. O en una encuesta de opinión es más relevante saber la proporción de opiniones favorables a cierta postura más que preocuparnos por el orden en que dichas opiniones fueron emitidas a lo largo del proceso de relevamiento.

Estas *consecuencias numéricas* pueden ser vistas como funciones, y dan lugar al concepto fundamental de *variable aleatoria*.

Definición 3.1 (variable aleatoria)

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. Una variable aleatoria es una función $X : \Omega \rightarrow \mathbb{R}$, satisfaciendo que

$$\{X \leq a\} = \{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{F}, \forall a \in \mathbb{R}. \quad (1)$$

Lo abreviamos diciendo que X es una v.a.

Ejemplo 3.1 (dos monedas)

Tiramos una moneda dos veces consecutivas

$\Omega = \{(S, S), (C, S), (S, C), (C, C)\}$ donde S representa una “ceca” y C representa una “cara”. Sea $X : \Omega \rightarrow \mathbb{R}$ dada por

$$X(\omega) = \#\text{caras obtenidas}$$

$$X(S, S) = 0, \quad X(C, S) = X(S, C) = 1, \quad X(C, C) = 2$$

Si un apostador gana \$1 por cada cara, la variable X representa la ganancia obtenida en el juego y nos interesará poder calcular la probabilidad de cada resultado: 0, 1, 2. Dependerá de la moneda. Si es equilibrada, $P(\omega) = 1/4 \quad \forall \omega \in \Omega$.

¿Cuándo X vale 0? La preimagen del $\{0\}$ por X ,

$$X^{-1}(\{0\}) = \{X = 0\} = \{\omega \in \Omega : X(\omega) = 0\} = \{(S, S)\}$$

En este caso, este conjunto es un evento elemental, y tenemos
 $P(\{X = 0\}) = P(\{\omega \in \Omega : X(\omega) = 0\}) = 1/4$

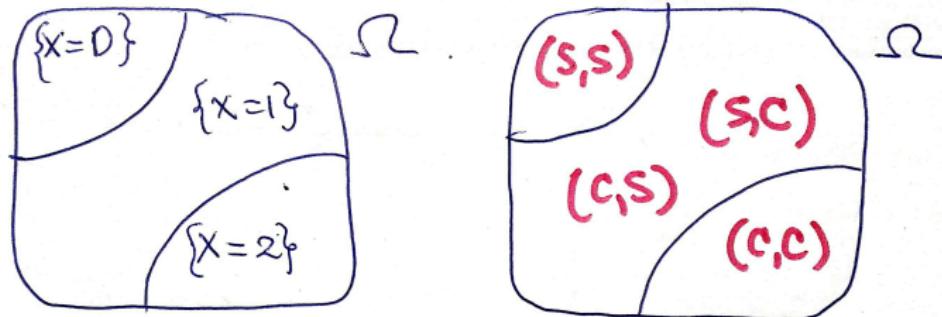
Ejemplo 3.1 (dos monedas, cont.)

$$P(\{X = 1\}) = P(\{(C, S), (S, C)\}) = 2/4$$

$$P(\{X = 2\}) = P(\{(C, C)\}) = 1/4$$

Por eso pedimos que estos conjuntos dados por las preimágenes de X estén en la σ -álgebra.

Observemos que esto parte al espacio muestral según el valor que toma X .



$$\Omega = \{x=0\} \cup \{x=1\} \cup \{x=2\} \quad (\text{disj.})$$

Ejemplo 3.2 (dardo)

Se elige un punto al azar sobre un tablero circular de radio siete :

$$\Omega = \{(\omega_1, \omega_2) \in \mathbb{R}^2 : \omega_1^2 + \omega_2^2 \leq 49\}.$$

La elección se hace al azar, de modo que

$$P(A) = \frac{\text{área}(A)}{\text{área tablero}} = \frac{\text{área}(A)}{\pi 49}.$$

Consideremos la variable aleatoria

$$X : \Omega \rightarrow \mathbb{R}$$

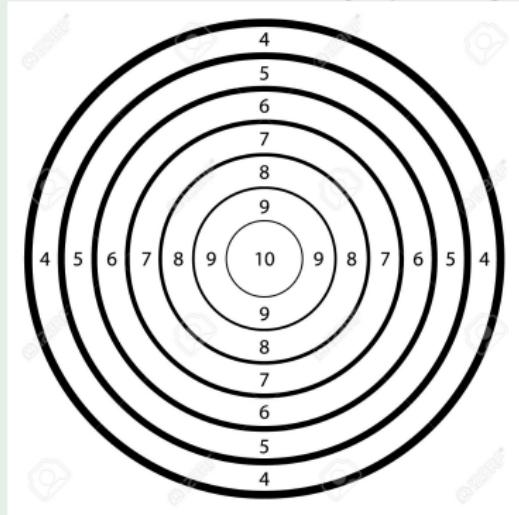
que a cada punto (ω_1, ω_2) le asigna su distancia al centro del tablero

$$X(\omega_1, \omega_2) = \sqrt{\omega_1^2 + \omega_2^2}.$$

Entonces X toma todos los valores comprendidos en el intervalo $[0, 7]$.

Ejemplo 3.2 (dardo, cont.)

Si pensamos que el tablero circular representa un blanco y que el punto elegido es el tiro del dardo, nuestro modelo impone que el dardo siempre se clava en el tablero, y que el jugador no tiene puntería especial.



Podríamos estar interesados en calcular la probabilidad de que el dardo caiga en la zona del círculo más cercano al centro. La región indicada con “10” en la figura, el círculo central de radio 1. Es decir, nos interesa

$$P(\{X \leq 1\}) = \frac{\pi}{49\pi} = \frac{1}{49}.$$

En ambos ejemplos, interesa asignarle probabilidad a eventos de la forma

$$\{X \leq a\}, \quad \{X = a\}, \quad \{a \leq X < b\}$$

Para que seamos capaces de calcular estas probabilidades, estos eventos deben pertenecer a \mathcal{F} . Esto justifica la definición de variable aleatoria que dimos.

Estos tipos de eventos serán el foco de nuestro interés.

Notación:

- En vez de escribir

$$X^{-1}(\{(-\infty, 1]\}) = \{\omega \in \Omega : X(\omega) \leq 1\} \quad \text{escribiremos} \quad \{X \leq 1\},$$

- En vez de $P(\{X \leq 1\})$ escribiremos $P(X \leq 1)$
- X (mayúscula del final del alfabeto) es la notación para las variables, mientras que las minúsculas (a, b, x) representarán los valores reales que toma

Función de distribución

Definición 3.2 (función de distibución)

Consideremos un espacio de probabilidad (Ω, \mathcal{F}, P) . Dada una variable aleatoria $X : \Omega \rightarrow \mathbb{R}$, definimos la función de distribución (acumulada) asociada a X , $F_X : \mathbb{R} \rightarrow [0, 1]$ mediante la fórmula

$$F_X(t) := P(X \leq t), \text{ para } t \in \mathbb{R}.$$

Calculemos las funciones de distribución de nuestros ejemplos (ejercicio).

Ejemplo 3.1 (dos monedas), función de distribución

$$F_X(t) = P(X \leq t).$$

Si $t < 0$, entonces $\{X \leq t\} = \{w \in \Omega : X(w) \leq t\}$.
y este conjunto es \emptyset , por lo que $F_X(t) = 0$ en
este caso.

Ejemplo 3.1 (dos monedas), función de distribución

$$F_X(t) = P(X \leq t).$$

Si $t < 0$, entonces $\{X \leq t\} = \{w \in \Omega : X(w) \leq t\}$.
y este conjunto es \emptyset , por lo que $F_X(t) = 0$ en
este caso.

$$\text{Si } t = 0, F_X(t) = P(X=0) = \frac{1}{4}$$

Observemos que $F_X(t) = \frac{1}{4} \quad \forall t \in [0, 1).$

Ejemplo 3.1 (dos monedas), función de distribución

$$\begin{aligned} \text{Si } t=1, F_X(1) &= P(X \leq 1) = P(\{X=0\} \cup \{X=1\}) \\ &= P(X=0) + P(X=1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \end{aligned}$$

Ejemplo 3.1 (dos monedas), función de distribución

$$\begin{aligned} \text{Si } t=1, \quad F_X(1) &= P(X \leq 1) = P(\{X=0\} \cup \{X=1\}) \\ &= P(X=0) + P(X=1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \end{aligned}$$

$$\text{Si } 1 \leq t < 2 \quad \Rightarrow F_X(t) = \frac{3}{4}.$$

Ejemplo 3.1 (dos monedas), función de distribución

$$\text{Si } t=0, F_X(t) = P(X=0) = \frac{1}{4}$$

Observemos que $F_X(t) = \frac{1}{4} \quad \forall t \in [0, 1]$.

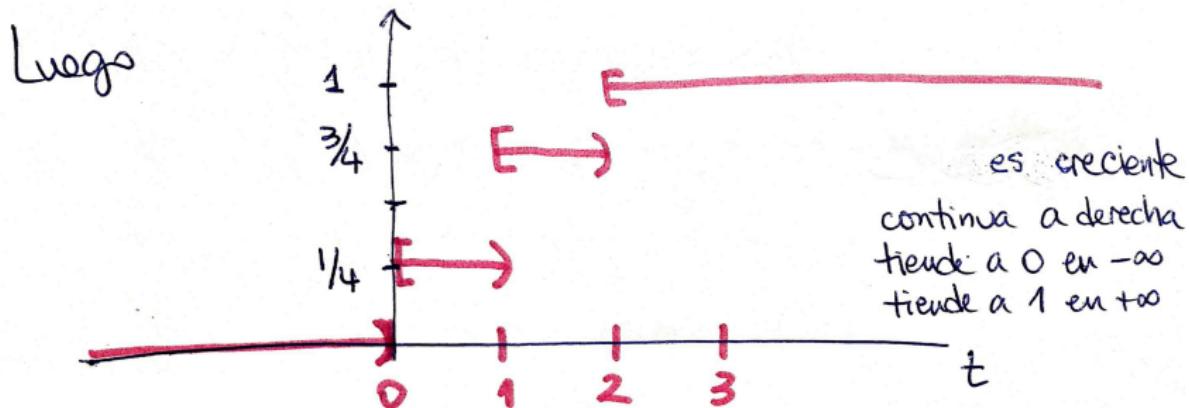
$$\begin{aligned}\text{Si } t=1, F_X(1) &= P(X \leq 1) = P(\{X=0\} \cup \{X=1\}) \\ &= P(X=0) + P(X=1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}\end{aligned}$$

$$\text{Si } 1 \leq t < 2 \Rightarrow F_X(t) = \frac{3}{4}.$$

$$\text{Si } t=2 \quad F_X(2) = P(X \leq 2) = P(\Omega) = 1.$$

$$F_X(t) = 1 \quad \forall t \geq 2.$$

Ejemplo 3.1 (dos monedas), gráfico de la función de distribución



Ejemplo 3.2, (dardo) función de distribución

Si $t < 0$, $F_X(t) = P(\underbrace{X \leq t}_{\text{distancia al origen}}) = P(\emptyset) = 0$.

distancia al origen
es menor a t que
es un N° negativo

$$\begin{aligned} \text{Si } 0 \leq t \leq 7, F_X(t) &= P(X \leq t) = \frac{\text{área (círculo de radio } t)}{\text{área (círculo de radio } 7)} \\ &= \frac{\pi t^2}{\pi 49} = \frac{t^2}{49} \end{aligned}$$

Ejemplo 3.2, (dardo) función de distribución

Si $t > 7$, $F_X(t) = P(X \leq t) = P(\Omega) = 1$

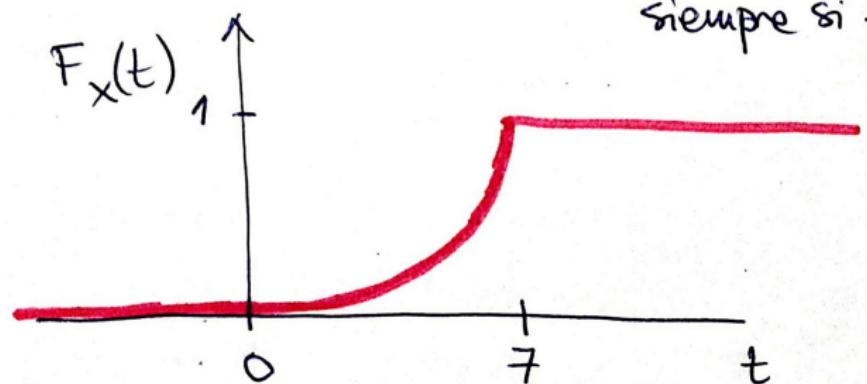
\uparrow
 $X \leq t$ ocurre
siempre si $t > 7$

Ejemplo 3.2, (dardo) función de distribución

$$\text{Si } t > 7, F_X(t) = P(X \leq t) = P(\Omega) = 1$$

\uparrow
 $X \leq t$ ocurre

\uparrow
siempre si $t > 7$



Es creciente, continua, $\lim_{t \rightarrow -\infty} F_X(t) = 0$ y $\lim_{t \rightarrow +\infty} F_X(t) = 1$.

Recordemos que valen las siguientes equivalencias, que serán útiles en la siguiente demostración.

- ① $\lim_{x \rightarrow x_0} g(x) = L \Leftrightarrow$ para toda sucesión $x_n \rightarrow x_0$ se tiene $\lim_{n \rightarrow \infty} g(x_n) = L$.
- ② $\lim_{x \searrow x_0} g(x) = L \Leftrightarrow$ para toda sucesión $x_n \rightarrow x_0$ y tal que $x_n > x_0$ resulta $\lim_{n \rightarrow \infty} g(x_n) = L \Leftrightarrow$ para toda sucesión x_n estrictamente decreciente y tal que $x_n \rightarrow x_0$ resulta $\lim_{n \rightarrow \infty} g(x_n) = L$.

Lema 3.3

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad y X una variable aleatoria $X : \Omega \rightarrow \mathbb{R}$, entonces, valen las siguientes propiedades de F_X ,

- ① $0 \leq F_X(t) \leq 1$ para todo $t \in \mathbb{R}$.
- ② F_X es una función creciente: si $s < t$, entonces $F_X(s) \leq F_X(t)$.
- ③ Como F_X es una función monótona y acotada, para todo $a \in \mathbb{R}$ existe el límite a izquierda que notaremos

$$F_X(a^-) := \lim_{t \rightarrow a^-} F_X(t) = \lim_{t \nearrow a} F_X(t)$$

.y el límite a derecha,

$$F_X(a^+) := \lim_{t \rightarrow a^+} F_X(t) = \lim_{t \searrow a} F_X(t).$$

Más aún, $F_X(a^+) = F_X(a)$ para todo $a \in \mathbb{R}$, es decir, F_X es continua a derecha.

- ④ $\lim_{t \rightarrow -\infty} F_X(t) = 0$ y $\lim_{t \rightarrow +\infty} F_X(t) = 1$.

Demostración.

① Vale por ser una probabilidad.

② Si $s < t$ entonces

$$(-\infty, s] \subset (-\infty, t], \text{ y por eso, } \{X \leq s\} \subset \{X \leq t\}$$

y por lo tanto

$$F_X(s) = P(X \leq s) \leq P(X \leq t) = F_X(t), \text{ (por monotonía)}$$



Demostración.

3. La existencia de los límites laterales es consecuencia de la monotonía.
Además

$$F_X(a^-) \leq F_X(a) \leq F_X(a^+)$$

Queremos ver que $\lim_{t \searrow a} F_X(t) = F_X(a)$. Sabemos que el límite existe. Sea t_n cualquier sucesión tal que $t_n \searrow a$ (por ejemplo, $t_n = a + \frac{1}{n}$), basta ver que vale $\lim_{n \rightarrow +\infty} F_X(t_n) = F_X(a)$. Pero, $F_X(t_n) = P(X \leq a + \frac{1}{n})$. La sucesión $\{X \leq (a + \frac{1}{n})\}, (n \in \mathbb{N})$ es una sucesión decreciente de conjuntos, y

$$\bigcap_{n \in \mathbb{N}} \left\{ X \leq \left(a + \frac{1}{n} \right) \right\} = \{X \leq a\}.$$

Luego, por continuidad de la probabilidad, tenemos

$$\lim_{n \rightarrow \infty} P \left(X \leq a + \frac{1}{n} \right) = P(X \leq a) = F_X(a).$$

Demostración.

4. Por ser F_X monótona y acotada, sabemos que ambos límites existen y que coinciden con los límites por sucesiones. Sea $\{X \leq n\}_{n \in \mathbb{N}}$, la sucesión de eventos $\{X \leq n\}$ es creciente y su unión es Ω , i.e. $\{X \leq n\} \nearrow \Omega$. Luego, por la continuidad de la probabilidad tenemos

$$\lim_{t \rightarrow +\infty} F_X(t) = \lim_{n \rightarrow +\infty} P(X \leq n) = P(\Omega) = 1.$$

El otro límite es análogo, tomando la sucesión $\{X \leq -n\} \searrow \emptyset$.



El Lema 3.3 caracteriza a las funciones de distribución

En realidad, estas 4 propiedades caracterizan a la función de distribución de una variable aleatoria, en el siguiente sentido

Lema 3.4

Si $F : \mathbb{R} \rightarrow [0, 1]$ y satisface las propiedades 2), 3) y 4) del Lema 3.3 entonces existe una variable aleatoria X definida en un espacio de probabilidad (Ω, \mathcal{F}, P) tal que $F_X = F$.

Veremos la demostración más adelante. De hecho, construiremos la variable aleatoria X y el espacio de probabilidad (Ω, \mathcal{F}, P) .

A partir de la función de distribución acumulada de la variable aleatoria X , podemos calcular

① $P(a < X \leq b) = P(X \in (-\infty, b] - (-\infty, a]) = F_X(b) - F_X(a).$

② $P(X < b) = F_X(b^-) = \lim_{t \rightarrow b^-} F_X(t).$ Luego
 $F_X(b^-) = F_X(b) - P(\{X = b\}).$

Dem: Sabemos que el límite que define a $F_X(b^-)$ existe, sea
 $t_n = b - \frac{1}{n}$, luego, como $\{X \leq t_n\} \nearrow \{X < b\}$, por la continuidad de la probabilidad tenemos

$$\begin{aligned}F_X(b^-) &= \lim_{t \rightarrow b^-} F_X(t) = \lim_{n \rightarrow \infty} F_X(t_n) = \lim_{n \rightarrow \infty} P\left(X \leq \left(b - \frac{1}{n}\right)\right) \\&= P(X < b) = P(X \leq b) - P(X = b) = F_X(b) - P(X = b)\end{aligned}$$

③ $P(X > a) = 1 - F_X(a).$

Ejercicio 3.1

- ① Expresar $P(a \leq X \leq b)$, $P(a \leq X < b)$, $P(a < X < b)$ usando F_X
- ② Una variable aleatoria X tiene función de distribución F . ¿Cuál es la función de distribución de $Y = aX + b$ donde a y b son constantes reales?
- ③ Muestre que si F y G son funciones de distribución y $0 \leq \lambda \leq 1$ entonces $\lambda F + (1 - \lambda)G$ es una función de distribución. ¿Es el producto FG una función de distribución? ¿Y el cuadrado de F , F^2 ?

Variables aleatorias y boreelianos

Proposición 3.1

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad y una variable aleatoria $X : \Omega \rightarrow \mathbb{R}$. Entonces, $X^{-1}(B) \in \mathcal{F}$ para todo $B \in \mathcal{B}(\mathbb{R}) = \mathcal{B}$.

Demostración.

Esta demostración, explota la condición de minimalidad que cumple la σ -álgebra generada por una colección de eventos. Sea

$$\mathcal{G} = \{B \in \mathcal{B}(\mathbb{R}) : X^{-1}(B) \in \mathcal{F}\}.$$

Primero: Veamos que \mathcal{G} es una σ -álgebra en \mathbb{R} .

- ① Queremos probar que $\mathbb{R} \in \mathcal{G}$. Observemos que $X^{-1}(\mathbb{R}) = \Omega$ pertenece a \mathcal{F} , luego $\mathbb{R} \in \mathcal{G}$.
- ② Es consecuencia de que la preimagen cumple $X^{-1}(B^c) = [X^{-1}(B)]^c$.



Demostración (cont.)

3. Sean $(B_i)_i \in \mathcal{G}$. Queremos probar que $\bigcup_{i=1}^{\infty} B_i \in \mathcal{G}$. Que $(B_i)_i \in \mathcal{G}$ quiere decir que los conjuntos $X^{-1}(B_i)$ pertenecen a \mathcal{F} para todo $i \in \mathbb{N}$. Como

$$\bigcup_{i=1}^{\infty} X^{-1}(B_i) = X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right)$$

y \mathcal{F} es cerrada bajo uniones numerables, resulta que $X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right) \in \mathcal{F}$ y por lo tanto $\bigcup_{i=1}^{\infty} B_i \in \mathcal{G}$.

Segundo: Veamos que $\mathcal{B}(\mathbb{R}) = \mathcal{G}$.

La colección de semirrectas, $\mathcal{I} = \{(-\infty, b] : b \in \mathbb{R}\} \subset \mathcal{G}$ por ser X una variable aleatoria. Luego $\sigma(\mathcal{I}) \subset \mathcal{G}$, por minimalidad de la σ -álgebra generada. Pero $\sigma(\mathcal{I}) = \mathcal{B}(\mathbb{R})$.

F_X caracteriza

Proposición 3.2

La función de distribución acumulada caracteriza a la variable aleatoria, en el sentido que si X e Y son variables aleatorias con $F_X = F_Y$, entonces

$$P(X \in A) = P(Y \in A), \quad \text{para todo } A \in \mathcal{B}(\mathbb{R}).$$

Sin demostración

F_X caracteriza

Observación 3.1

¿Cómo debe interpretarse la Proposición 3.2? Consideremos los dos experimentos aleatorios siguientes:

- ① **Experimento 1.** Tiramos 5 veces un dado, de forma sucesiva. Sea $X = \text{cantidad de resultados pares obtenidos}$.
- ② **Experimento 2** Tiramos una moneda 5 veces de manera consecutiva. Sea $Y = \text{cantidad de caras obtenidas}$.

Los espacios muestrales para estos experimentos son:

$$\Omega_1 = \{(a_1, a_2, \dots, a_5) : \text{con } 1 \leq a_i \leq 6\}$$

$$\Omega_2 = \{(b_1, b_2, \dots, b_5) : \text{con } b_i \in \{\text{cara, ceca}\}\}$$

Observemos que $X \neq Y$. Sin embargo, $F_X = F_Y$ y $P(X \in A) = P(Y \in A)$ para todo $A \in \mathcal{B}(\mathbb{R})$.

Variables Aleatorias Discretas

Definición 3.3 (v.a. discreta)

Diremos que una variable aleatoria es discreta cuando existe un conjunto $A \subset \mathbb{R}$ finito o numerable de forma tal que

$$P(X \in A) = 1. \quad (2)$$

A no está únicamente determinado. Se le pueden agregar puntos a con $P(X = a) = 0$. Para una variable aleatoria discreta, es relevante entonces caracterizar el menor conjunto A para el cual se verifica la propiedad (2).

Definición 3.4 (rango)

El rango R_X de una variable aleatoria discreta X está dado por el conjunto de puntos para los cuales la *preimagen tiene probabilidad diferente de cero*,

$$R_X = \{k_i \in \mathbb{R} : P(X = k_i) \neq 0\}.$$

Función de probabilidad puntual

Definición 3.5 (función de probabilidad puntual)

Dada una variable aleatoria discreta X , definimos la función de probabilidad puntual asociada a X mediante la fórmula, $p_X : R_X \rightarrow [0, 1]$

$$p_X(k_i) := P(X = k_i) = F_X(k_i) - F_X(k_i^-). \quad (3)$$

Ejemplo: Volvamos al ejemplo de tirar 1 moneda ^{2 veces,} sucesivamente y sea $X = \#$ caras obtenidas. X es discreta

pues su imagen

se concentra en un conjunto numerable,

$$R_X = \{0, 1, 2\}$$

pues

$$P(X \in \{0, 1, 2\}) = 1.$$

$$\text{y } P_X(0) = P(CC) = 1/4$$

$$P_X(1) = \frac{1}{2}$$

$$P_X(2) = 1/4.$$

Podemos graficar la función de probabilidad puntual mediante un diagrama de barras

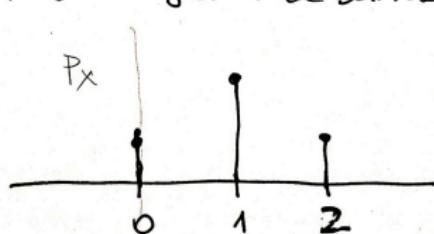
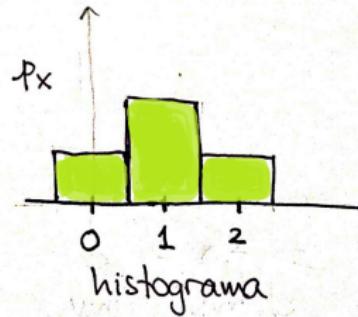


diagrama de barras



histograma

Lema 3.5

Sea X una variable aleatoria discreta con rango $R_X = \{k_i : i \geq 1\}$.

Tenemos entonces que

- ① $P(X \in R_X) = 1$
- ② $P(X \in B) = P(X \in B \cap R_X) = \sum_{k_i \in B} P(X = k_i)$ para todo B boreliano
- ③ $F_X(t) = \sum_{k_i \in R_X : k_i \leq t} P(X = k_i)$

Demostración.

1. Como X es v.a. discreta existe A finito o numerable tal que $P(X \in A) = 1$. Por probabilidad total, tenemos

$$P(X \in R_X) = P(X \in R_X \cap A) + P(X \in R_X \cap A^c) \quad (4)$$

Como $\{X \in R_X \cap A^c\} \subset \{X \in A^c\}$ y $\Omega = \{X \in A\} \cup \{X \in A^c\}$ entonces $1 = P(\Omega) = P(X \in A) + P(X \in A^c)$ resulta $P(X \in A^c) = 0$ y por monotonía $P(X \in R_X \cap A^c) = 0$. Reemplazando en (4) tenemos

$$P(X \in R_X) = P(X \in R_X \cap A) \quad (5)$$

Por probabilidad total tenemos

$$P(X \in A) = P(X \in A \cap R_X) + P(X \in A \cap R_X^c) \quad (6)$$

Por definición de R_X , si $b \in R_X^c$ resulta $P(X = b) = 0$. Como $A \cap R_X^c$ es a lo sumo numerable (por estar contenido en A) resulta $P(X \in A \cap R_X^c) =$

$$P\left(X \in \bigcup_{b \in A \cap R_X^c} \{b\}\right) = P\left(\bigcup_{b \in A \cap R_X^c} \{X = b\}\right) = \sum_{b \in A \cap R_X^c} P(X = b) = 0.$$

Reemplazando en (6) tenemos

$$1 = P(X \in A) = P(X \in A \cap R_X) \underbrace{=}_{\text{por (5)}} P(X \in R_X). \text{ Observemos que de esto se}$$

deduce que $P(X \in R_X^c) = 1 - P(X \in R_X) = 0$.

Demostración.

2. $P(X \in B) = P(X \in B \cap R_X) + P(X \in B \cap R_X^c)$ usando probabilidad total. Como $\{X \in B \cap R_X^c\} \subset \{X \in R_X^c\}$ que es un evento de probabilidad cero, resulta, por monotonía $P(X \in B) = P(X \in B \cap R_X) = P\left(X \in \bigcup_{k_i \in B} \{k_i\}\right) = \sum_{k_i \in B} p_X(k_i)$.
3. Caso particular del ítem 2), tomando $B = (-\infty, t]$



Variables discretas: relación entre p_X y F_X

Podemos expresar a F_X en términos de p_X , usando el ítem 3) del lema anterior:

$$F_X(t) = \sum_{k_i \in R_X : k_i \leq t} p_X(k_i)$$

Por lo que en el caso de X discreta, F_X es constante por intervalos y salta en aquellos puntos donde p_X es distinta de cero (i.e. los puntos del R_X) Al revés: ¿cómo expresamos a p_X si conocemos a F_X ?

$$p_X(k) = P(X = k) = F_X(k) - F_X(k^-)$$

Pasamos de p_X \rightarrow F_X sumando términos

Pasamos de F_X \rightarrow p_X restando términos

Para poder calcular probabilidades que involucren a X basta

conocer p_X o F_X Nos referiremos a la distribución de X tanto si hablamos de la función de probabilidad puntual o de la función de distribución de X

Función indicadora

Definición 3.6

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. Dado $A \subseteq \Omega$, definimos $I_A : \Omega \rightarrow \mathbb{R}$, la función indicadora del conjunto A mediante

$$I_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A \\ 0 & \text{caso contrario} \end{cases}$$

Ejercicio 3.2

Demuestre que I_A es una variable aleatoria si y sólo si $A \in \mathcal{F}$.

Ejercicio 3.3

Sea $\{B_j : j \in J\}$ una familia de eventos disjuntos tales que $A \subset \bigcup_{j \in J} B_j$. Entonces $I_A = \sum_{j \in J} I_{A \cap B_j}$.

Distribuciones discretas famosas

Bernoulli

Asociado a un experimento con dos posibles resultados, generalmente identificados como éxito y fracaso en un espacio (Ω, \mathcal{F}, P) puede definirse la variable aleatoria $X : \Omega \rightarrow \mathbb{R}$ que toma los valores

$X = 1$ si observamos éxito, y

$X = 0$ caso contrario.

La función de probabilidad puntual está dada por

$$p_X(0) = P(X = 0) = 1 - p$$

$$p_X(1) = P(X = 1) = p$$

siendo p la probabilidad de observar un éxito, con $0 \leq p \leq 1$. En este caso decimos que X es una variable aleatoria con *distribución Bernoulli* de parámetro p

y lo notaremos $X \sim Be(p)$. El símbolo \sim en el contexto probabilístico reemplaza la frase “se distribuye como”. Al experimento que produce dos resultados se lo suele llamar *ensayo Bernoulli*.

Distribución Binomial

Suponga ahora que realizamos n repeticiones independientes de un ensayo Bernoulli y que en cada repetición podemos observar éxito con probabilidad p y fracaso con probabilidad $1 - p$. Sea X la variable aleatoria $X = \text{cantidad de éxitos obtenidos en las } n \text{ repeticiones}$.

Entonces X es una variable aleatoria discreta con rango $0, 1, \dots, n$.

(*¿Cuál es Ω ? ¿Es equiprobable? ¡Pensarlo!*)

Su función de probabilidad puntual está caracterizada por los parámetros (n, p) y viene dada por

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \text{ para } 0 \leq k \leq n \quad (7)$$

Diremos que la variable

X tiene *distribución binomial* con parámetros (n, p) y lo notaremos:
 $X \sim \mathcal{B}(n, p)$.

Ejercicio: Ω para Binomial

Ejercicio 3.4

Repetimos un ensayo Bernoulli de parámetro p , n veces de forma independiente. Sea

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in \{0, 1\}\}.$$

$\omega_i = 1 \Leftrightarrow$ la i -ésima repetición fue un éxito.

- a) Calcular $\#\Omega$.
- b) $P(\omega) = \prod_{i=1}^n p^{\omega_i} (1-p)^{1-\omega_i} = p^{\sum_{i=1}^n \omega_i} (1-p)^{\sum_{i=1}^n (1-\omega_i)}$, verificarlo.
¿Para qué valores de p resulta Ω equiprobable?
- c) Sea $X = \#$ de éxitos en n intentos. Observar que
 $P(\omega) = p^k (1-p)^{n-k}$ para todo $\omega \in \{X = k\}$.
- d) $\{X = k\} = \bigcup_{\omega \in \{X = k\}} \{\omega\}$. Calcular el cardinal de $\{X = k\}$ y comprobar que (7) es la expresión correcta para la función de probabilidad puntual de X .

Ejemplo Binomial

El examen final de Matemática del CBC es multiple choice. Son 20 ejercicios con 4 opciones cada uno (sólo una opción es correcta). Para aprobar el examen hay que tener al menos 8 respuestas correctas y más respuestas correctas que incorrectas. Es vox populi el siguiente consejo a los alumnos:

“contestar 15 preguntas solamente, pues esto aumenta las chances de aprobar el final”. ¿Será cierto? Por supuesto, si uno supiera las respuestas de todas las preguntas, lo más sensato es contestar todo el examen. Así que asumimos que el alumno no sabe las respuestas, y **contesta al azar todas las preguntas, de forma independiente.** Inicialmente comparemos dos escenarios:

- A Contesta 15 preguntas
- B Contesta 20 preguntas

¿Con cuál de los dos escenarios es mayor la probabilidad de aprobar?

Estrategia A

Sea $S_{15} = \#$ de preguntas bien contestadas al contestar 15 preguntas.

$S_{15} \sim \mathcal{B}(n = 15, p = P(\text{éxito}) = 1/4)$.

$$P(\text{aprobar en el escenario A}) = P(S_{15} \geq 8 \text{ y } S_{15} > 15 - S_{15})$$

$$= P(\{S_{15} \geq 8\} \cap \{S_{15} > 7,5\}) = P(S_{15} \geq 8) = \sum_{k=8}^{15} p_{S_{15}}(k)$$

$$= 1 - P(S_{15} \leq 7) = 1 - F_{S_{15}}(7) = 0,01729 \quad \text{¡baja!}$$

En R:

```
> probaprobar<-1-pbinom(7, size = 15,prob = 0.25)
> probaprobar
[1] 0.01729984
```

`pbinom` nos da la función de distribución `dbinom` nos da la función de probabilidad puntual, o sea que también la podemos calcular con

```
sum(dbinom(8:15, size = 15,prob = 0.25 ))
```

Estrategia B

Sea $S_{20} = \#$ de preguntas bien contestadas al contestar 20 preguntas.
 $S_{20} \sim \mathcal{B}(n = 20, p = P(\text{éxito}) = 1/4)$.

$$\begin{aligned}P(\text{aprobar en el escenario A}) &= P(S_{20} \geq 8 \text{ y } S_{20} > 20 - S_{20}) \\&= P(\{S_{20} \geq 8\} \cap \{S_{20} > 10\}) = P(S_{20} > 10) = \sum_{k=11}^{20} p_{S_{20}}(k) \\&= 1 - P(S_{20} \leq 10) = 1 - F_{S_{20}}(10) = 0,003942 \quad \text{¡más baja!}\end{aligned}$$

Para darnos cuenta cuánto más baja es, calculemos

$$\frac{P(\text{aprobar con estrategia A})}{P(\text{aprobar con estrategia B})} = 4,388$$

Estrategias intermedias, ¿habrá una mejor?

Sea $S_n = \#$ de preguntas bien contestadas al contestar n preguntas.
 $8 \leq n \leq 20$ $S_n \sim \mathcal{B}(n, p = P(\text{éxito}) = 1/4)$.

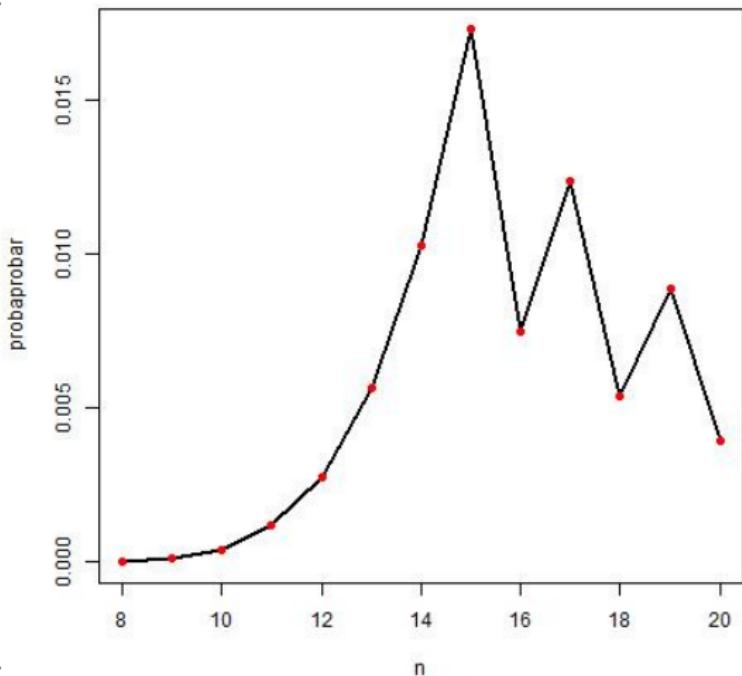
$P(\text{aprobar con la estrategia de contestar } n \text{ preguntas})$

$$\begin{aligned} &= P(S_n \geq 8 \text{ y } S_n > n - S_n) = P(\{S_n \geq 8\} \cap \{S_n > \frac{n}{2}\}) \\ &= P(S_n > \max\{7, n/2\}) = 1 - F_{S_n}(\max\{7, n/2\}) \end{aligned}$$

Calculamos con R: probabilidad de aprobar

n	Proba Aprobar
-----	---------------

8	0.00002
9	0.0001
10	0.0004
11	0.0012
12	0.0028
13	0.0056
14	0.0103
15	0.0173
16	0.0075
17	0.0124
18	0.0054
19	0.0089
20	0.0039

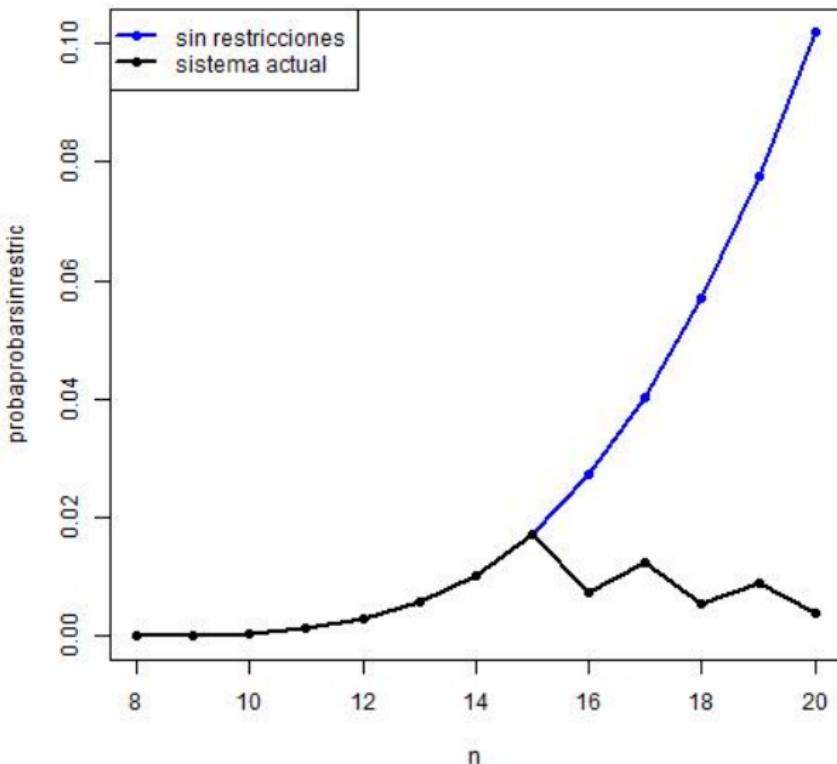


$n \quad 1 - F_{S_n}(\max\{7, n/2\})$

Instrucciones en R

```
# n=15
probaprobar<-1-pbinom(7, size = 15,prob = 0.25)
probaprobar
sum(dbinom(8:15, size = 15,prob = 0.25 ))  
  
# n=20
probaprobar<-1-pbinom(10, size = 20,prob = 0.25)
probaprobar  
  
# caso general
ene<-8:20  
  
probaprobar<-1-pbinom(pmax(7,ene/2), size = ene,prob = 0.25)  
  
plot(ene,probaprobar,type="l",lwd=2,xlab="n")
points(ene,probaprobar,pch=16,col="red")
```

¿Y si no hubiera restricciones de # respuestas incorrectas?



¿Y si no hubiera restricciones de # respuestas incorrectas?

```
# si no hubiera restricciones sobre cantidad de respuestas
# incorrectas (instrucciones en R)
probaprobarsinrestric<-1-pbinom(7, size = ene,prob = 0.25)
plot(ene,probaprobarsinrestric,type="l",lwd=2,col="blue"
,xlab = "n")
points(ene,probaprobarsinrestric,col="blue",pch=16)

points(ene,probaprobar,pch=16,col="black")
lines(ene,probaprobar,lwd=2)
max(probaprobarsinrestric)
legend("topleft",c("sin restricciones","sistema actual"),
pch=c(16,16),lwd=c(2,2),col=c("blue","black"))
```

¿Seguirá siendo cierto el consejo si $p = 0,9$? ¿Y si $p = 0,5$?

Distribución Hipergeométrica

Consideremos una población de tamaño N con M éxitos. Extraemos una muestra de tamaño n , **sin reposición**. Nos interesa la v.a.

X = número de éxitos en la muestra .

X es una v.a. discreta, su distribución dependerá de 3 parámetros:

N = tamaño de la población,

M = número de éxitos en la población,

n = tamaño de la muestra.

(*¿Cuál es Ω ? ¿Es equiprobable?*) La probabilidad puntual es:

$$p_X(k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \quad (8)$$

en el denominador tenemos todas las formas posibles de extraer la muestra y en el numerador tenemos la cantidad de muestras con exactamente k éxitos (y entonces $n - k$ fracasos).

Distribución Hipergeométrica

¿Para qué valores de k tiene sentido la fórmula de p_X ?

$$P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$$

$0 \leq k \leq M$ y $0 \leq n - k \leq N - M$. O resumiendo, el rango de X está dado por

$$\max\{0, n - N + M\} \leq k \leq \min\{n, M\}.$$

Diremos que la variable aleatoria X tiene *distribución hipergeométrica* con parámetros N, M, n y lo notaremos: $X \sim H(N, M, n)$.

Ejemplo 3.6

Consideremos una urna con 5 bolillas blancas y 7 bolillas rojas. Se extraen 3 bolillas sin reposición. Sea X el número de bolillas rojas extraídas.

Entonces $X \sim H(N = 12, M = 7, n = 3)$.

Relación entre binomial e hipergeométrica

Si la muestra hubiese sido elegida con reemplazo, y definimos Y el número de éxitos al extraer una muestra de tamaño n con reemplazo. Entonces $Y \sim \mathcal{B}(n, p = \frac{M}{N})$. Cuando el tamaño de la población es muy grande comparado con el tamaño de la muestra, podemos pensar que sacar con o sin reemplazo no debería alterar mucho la situación.

Podemos entonces aproximar una hipergeométrica por una distribución Binomial obteniendo que

$$H(N, M, n) \approx \mathcal{B}(n, M/N) . \quad (9)$$

Utilizaremos este tipo de aproximaciones siempre que $n/N \leq 0,05$ y siempre que M/N no esté muy cerca ni de cero ni de uno. Precisaremos esta aproximación más adelante.

Distribución Geométrica

Tenemos un ensayo Bernoulli, es decir, un experimento cuyos posibles resultados son éxito o fracaso. Repetimos ese experimento de forma independiente. Sea X la variable que cuenta el número de repeticiones hasta obtener el primer éxito. X es discreta. Si p es la probabilidad de éxito en cada repetición, tenemos entonces que

$$P(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, \dots$$

Decimos que X tiene distribución Geométrica con parámetro p y notamos: $X \sim \mathcal{G}(p)$.

Ejercicio 3.5

Verificar que $\sum_{k=1}^{\infty} p_X(k) = 1$. Hallar la función de distribución de X .

Proposición 3.3 (Falta de Memoria discreta)

Sea $X \sim \mathcal{G}(p)$, para todo $n, m \in \mathbb{N}$ tenemos que

$$P(X \geq n + m \mid X > m) = P(X \geq n).$$

(ejercicio de la práctica)

Distribución Binomial Negativa

Repetimos de forma independiente un ensayo Bernoulli con probabilidad p de éxito en cada repetición. Definimos la variable aleatoria

X = número de repetición en el que ocurre el r -ésimo éxito.

$R_X = \{r, r+1, \dots\}$. El evento $\{X = k\}$ está conformado por todas las tiras (uplas) de longitud k que contienen exactamente r éxitos y $k - r$ fracasos. El último éxito ocurre en el último intento en estas tiras.

- ① Fijada una upla ω en $\{X = k\}$, ¿cómo calculamos la probabilidad?
Por independencia, $P(\omega) = p^r(1 - p)^{k-r}$
¡todas tienen igual probabilidad!
- ② ¿Cuántas k -uplas hay en $\{X = k\}$? Tantas como formas distintas de elegir los $r - 1$ lugares donde se ubicarán los restantes $(k - 1)$ éxitos.

Finalmente, poniendo todo junto, podemos calcular la función de probabilidad puntual

$$P(X = k) = P\left(\bigcup_{\omega \in \{X=k\}} \{\omega\}\right) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \text{ para } k \geq r.$$

Distribución Binomial Negativa

En este caso decimos que X tiene una *distribución Binomial Negativa* con parámetros (r, p) y lo notamos por: $X \sim BN(r, p)$.

Ejercicio 3.6

Dos tenistas A y B igualmente hábiles juegan un partido al mejor de 5 sets. ¿Cuál es la probabilidad de que el partido termine al finalizar el cuarto set?

Distribución Poisson

La función de probabilidad puntual para una variable aleatoria con *distribución Poisson* de parámetro $\lambda > 0$, lo cual notaremos $X \sim \mathcal{P}(\lambda)$, es

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, k \in \mathbb{N}_0.$$

Por el desarrollo en potencias de la función exponencial, podemos garantizar que las p_X así definidas suman uno. La distribución de Poisson aparece naturalmente como límite para distribuciones binomiales $Y_n \sim \mathcal{B}(n, p_n)$, cuando $n \rightarrow \infty$, y $np_n \rightarrow \lambda$. Más específicamente, tenemos

Lema 3.7 (aproximación Poisson a la Binomial)

Sea $\lambda > 0$ y $(p_n)_{n \geq 1}$ una sucesión de números en $[0, 1]$ tal que $np_n \rightarrow \lambda$.
Sea $Y_n \sim \mathcal{B}(n, p_n)$, entonces

$$\lim_{n \rightarrow \infty} p_{Y_n}(k) = p_X(k), \text{ para todo } k \in \mathbb{N}_0,$$

con $X \sim \mathcal{P}(\lambda)$.

Distribución Poisson

Demostración.

Sabemos que $np_n \rightarrow \lambda$. Esto implica que $p_n \rightarrow 0$.

$$p_{Y_n}(k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k}$$

$$= \overbrace{\frac{n(n-1)\cdots(n-k+1)}{k!}}^{k \text{ factores}} p_n^k (1 - p_n)^n (1 - p_n)^{-k}$$

$$\lim_{n \rightarrow \infty} (1 - p_n)^n = \lim_{n \rightarrow \infty} \left[(1 + (-p_n))^{-\frac{1}{p_n}} \right]^{-np_n} = e^{-\lambda}$$

$$\lim_{n \rightarrow \infty} n(n-1)\cdots(n-k+1)p_n^k = \lim_{n \rightarrow \infty} \prod_{i=0}^{k-1} [(n-i)p_n] = \lambda^k$$

$$\lim_{n \rightarrow \infty} (1 - p_n)^{-k} = 1$$

Juntando todo tenemos

$$\lim_{n \rightarrow \infty} p_{Y_n}(k) = \lambda^k e^{-\lambda} \frac{1}{k!} = p_X(k), \quad X \sim \mathcal{P}(\lambda)$$

Distribución Poisson

La distribución Poisson suele emplearse para modelar la cantidad de éxitos que ocurren en una situación que puede pensarse que hay una cantidad muy grande de ensayos Bernoulli y la probabilidad de éxito en cada uno es muy pequeña.

Podemos pensar de forma intuitiva que si tenemos un fenómeno poco frecuente (por ejemplo, un incendio) que ocurre a lo largo del tiempo (por ejemplo, un año) podemos subdividir el tiempo en n subintervalos $(t_i, t_i + \Delta_n]$ tan cortos como para que en cada subintervalo pueda ocurrir a lo sumo un éxito (un incendio) de modo que podemos modelar adecuadamente la distribución de la variable aleatoria

$X =$ cantidad de incendios ocurridos en una cierta ciudad en un año

como el límite de distribuciones binomiales (que cuentan la cantidad de éxitos (incendios) en n intentos), es decir, con una distribución Poisson.

Distribución Poisson

En los ejercicios prácticos

se informará que la variable de interés tiene distribución Poisson y con qué valor de λ hay que trabajar. Ejemplos de variables con distribución Poisson:

- ① La cantidad de reclamos de un tipo (por ejemplo, robo de hogares) que recibe una compañía de seguros durante un año.
- ② La cantidad de errores de tipeo en una página de texto publicada.
- ③ La cantidad de pasajeros con ticket de vuelo comprado en una aerolínea que no se presenta al vuelo, en un mes.
- ④ La cantidad de partículas emitidas por un determinado volumen de cierto material radiactivo en una hora.

4. Variables Aleatorias Continuas

Probabilidades y Estadística (M)

María Eugenia Szretter Noste

Departamento de Matemática e
Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Primer cuatrimestre 2020



4. Variables Aleatorias (absolutamente) Continuas

Estamos interesados en elegir al azar un punto en el intervalo $(0, 2]$. A diferencia de los ejemplos considerados hasta ahora, el conjunto de posibles valores del experimento es no numerable. ¿Cómo hacemos entonces para calcular probabilidades? Si denotamos por X la variable aleatoria que indica el punto elegido, es razonable asumir que la probabilidad de que $X \in (a, b]$ debería ser proporcional a la longitud del intervalo, para todo $[a, b] \subset (0, 2]$. Es decir:

$$P(X \in (a, b]) = C(b - a), \text{ para todo } 0 < a \leq b \leq 2.$$

¿Cuál es el valor de la constante C ? Siendo que $P(X \in (0, 2]) = 1$, tenemos que

$$C(2 - 0) = 1,$$

de donde deducimos que $C = 1/2$. Tenemos entonces que

$$P(X \in (a, b]) = [1/2](b - a), \text{ para todo } 0 < a \leq b \leq 2. \quad \text{¿Cómo}$$

calcular la función de distribución asociada a la variable aleatoria X ?

Ejemplo, elegir un punto al azar

$$P(X \in (a, b]) = [1/2](b - a), \text{ para todo } 0 < a \leq b \leq 2.$$

¿Cómo

calcular la función de distribución asociada a la variable aleatoria X ?

Recordemos que

$$F_X(x) = P(X \leq x).$$

Así, tenemos que si $x \leq 0$, $F_X(x) = 0$. Si $x \in [0, 2)$, $F_X(x) = 1/2x$ y finalmente, la función de distribución vale 1 a partir de $x = 2$:

$$F_X(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ \frac{x}{2} & \text{si } 0 < x \leq 2, \\ 1 & \text{si } x > 2. \end{cases}$$

Obsérvese que si definimos la función $f_X(x)$ mediante la formula

$$f_X(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ \frac{1}{2} & \text{si } 0 < x \leq 2, \\ 0 & \text{si } x > 2, \end{cases}$$

Tenemos que $F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(u)du$.

Ejemplo, elegir un punto al azar

Tenemos que $F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(u) du$.

Es decir, integrando la función f_X desde menos infinito hasta x obtenemos la función de distribución F_X evaluada en x . En tal caso, diremos que **f_X es la función de densidad** de la variable aleatoria X . ¿Qué propiedades debe satisfacer una densidad? Simple: ser positiva e integrar uno en toda la recta.

Definición 4.1 (variable aleatoria absolutamente continua)

Una variable aleatoria X se dice **absolutamente continua** si existe $f_X : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ tal que

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(u) du .$$

En tal caso, diremos que f_X es la **función de densidad** de la variable aleatoria X .

Función de densidad

Observación 4.1

Si X tiene función de densidad f_X , tenemos que

$$1 = \lim_{x \rightarrow \infty} F_X(x) = \lim_{x \rightarrow \infty} \int_{-\infty}^x f_X(u) du = \int_{-\infty}^{+\infty} f_X(u) du ,$$

de donde concluimos que

$$\int_{-\infty}^{+\infty} f_X(u) du = 1 .$$

Definición 4.2 (función de densidad)

*Toda función $f \geq 0$ tal que $\int_{-\infty}^{+\infty} f(u) du = 1$ se dice **función de densidad**.*

Variable aleatoria continua

Definición 4.3 (Variable aleatoria continua)

La variable aleatoria X se dice **continua** si su función de distribución acumulada $F_X : \mathbb{R} \rightarrow [0, 1]$ es continua. En otras palabras, X se dice continua si $P(X = b) = 0$ para todo $b \in \mathbb{R}$.

¿Por qué? Porque habíamos visto que

$$P(X < b) = F_X(b^-) = \lim_{t \rightarrow b^-} F_X(t). \text{ Luego}$$

$$F_X(b^-) = F_X(b) - P(X = b). \text{ Si } F_X \text{ es continua, entonces}$$

$$F_X(b^-) = F_X(b) \quad \forall b \in \mathbb{R}, \text{ y por lo tanto } P(X = b) = 0$$

Lema 4.1

Si X es absolutamente continua, entonces X es continua.

Demostración.

Qvq F_X es continua en x_0 .

- Si f_X es acotada en un entorno de x_0 , $|f_X(u)| \leq c$ entonces $|F_X(x_0) - F_X(x)| \leq c|x_0 - x| \rightarrow 0$ cuando $x \rightarrow x_0$.
- Si f_X no es acotada en un entorno de x_0 , entonces $F_X(x_0) = \int_{-\infty}^{x_0} f_X(u) du$ se define como integral impropia,

$$F_X(x_0) = \int_{-\infty}^{x_0} f_X(u) du = \lim_{x \rightarrow x_0} \int_{-\infty}^x f_X(u) du = \lim_{x \rightarrow x_0} F_X(x).$$

Luego, F_X es continua en x_0 .



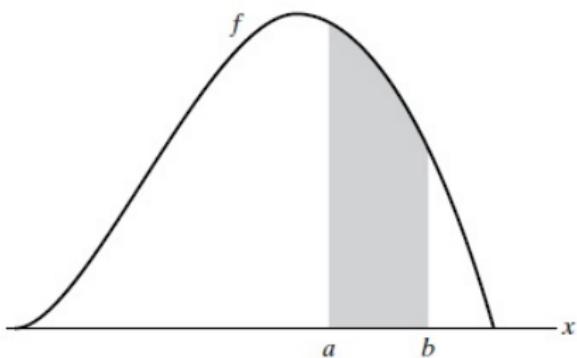
Recordemos que, conociendo la función de distribución de una variable aleatoria, podíamos calcular varias probabilidades. Por ejemplo, teníamos

que $P(X \in (a, b]) = P(a < X \leq b) = F_X(b) - F_X(a)$.

Como la distribución se obtiene de integrar la densidad, tenemos que

$$P(a < X \leq b) = F_X(b) - F_X(a) =$$

$$\int_{-\infty}^b f_X(u) du - \int_{-\infty}^a f_X(u) du = \int_a^b f_X(u) du.$$



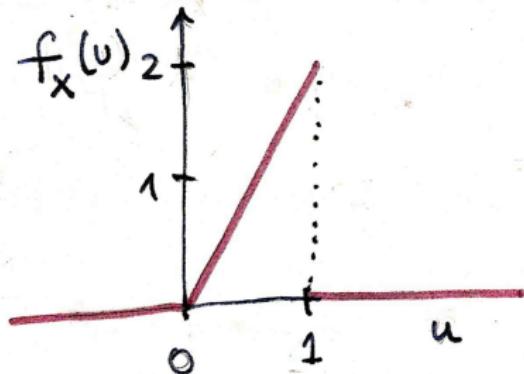
$$P(a \leq X \leq b) = \text{area of shaded region}$$

¿Qué (no) es la densidad?

Observación 4.2

f_X no es una probabilidad. De hecho, puede tomar valores mayores a 1.

Por ejemplo $f_X(u) = \begin{cases} 2u & \text{si } u \in (0,1) \\ 0 & \text{en caso contrario} \end{cases}$



$$= 2u I_{(0,1)}(u)$$

$$f_X(0.9) = 1.8$$

Lema 4.2

Si f_X es una función de densidad continua en x_0 , entonces

$$\lim_{h \rightarrow 0} \frac{P(X \in [x_0 - h, x_0 + h])}{2h} = f_X(x_0).$$

Lema 4.2

Si f_X es una función de densidad continua en x_0 , entonces

$$\lim_{h \rightarrow 0} \frac{P(X \in [x_0 - h, x_0 + h])}{2h} = f_X(x_0).$$

dem: Sean $m_h = \min \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

$M_h = \max \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

Lema 4.2

Si f_X es una función de densidad continua en x_0 , entonces

$$\lim_{h \rightarrow 0} \frac{P(X \in [x_0 - h, x_0 + h])}{2h} = f_X(x_0).$$

dem: Sean $m_h = \min \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

$M_h = \max \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

Por continuidad, $\lim_{h \rightarrow 0} m_h = \lim_{h \rightarrow 0} M_h = f_X(x_0)$

Lema 4.2

Si f_X es una función de densidad continua en x_0 , entonces

$$\lim_{h \rightarrow 0} \frac{P(X \in [x_0 - h, x_0 + h])}{2h} = f_X(x_0).$$

dem: Sean $m_h = \min \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

$M_h = \max \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

Por continuidad, $\lim_{h \rightarrow 0} m_h = \lim_{h \rightarrow 0} M_h = f_X(x_0)$

$$2h m_h \leq \int_{x_0 - h}^{x_0 + h} f_X(u) du \leq 2h M_h$$

Lema 4.2

Si f_X es una función de densidad continua en x_0 , entonces

$$\lim_{h \rightarrow 0} \frac{P(X \in [x_0 - h, x_0 + h])}{2h} = f_X(x_0).$$

dem: Sean $m_h = \min \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

$M_h = \max \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

Por continuidad, $\lim_{h \rightarrow 0} m_h = \lim_{h \rightarrow 0} M_h = f_X(x_0)$

$$2h m_h \leq \int_{x_0 - h}^{x_0 + h} f_X(u) du \leq 2h M_h$$

$$\text{O por, } m_h \leq \frac{1}{2h} \int_{x_0 - h}^{x_0 + h} f_X(u) du \leq M_h$$

Lema 4.2

Si f_X es una función de densidad continua en x_0 , entonces

$$\lim_{h \rightarrow 0} \frac{P(X \in [x_0 - h, x_0 + h])}{2h} = f_X(x_0).$$

dem: Sean $m_h = \min \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

$M_h = \max \{f_X(u) : u \in [x_0 - h, x_0 + h]\}$.

Por continuidad, $\lim_{h \rightarrow 0} m_h = \lim_{h \rightarrow 0} M_h = f_X(x_0)$

$$2h m_h \leq \int_{x_0 - h}^{x_0 + h} f_X(u) du \leq 2h M_h$$

O por, $m_h \leq \frac{1}{2h} \int_{x_0 - h}^{x_0 + h} f_X(u) du \leq M_h$

El resultado se obtiene pasando al límite cuando $h \rightarrow 0$.

Teorema 4.3

Si f_X es continua en x_0 entonces, F_X es derivable en x_0 y además

$$F'_X(x_0) = f_X(x_0).$$

Demostración.

Vale por el Teorema fundamental del Cálculo



Observación 4.3

Observemos que si modificamos la definición de la densidad f_X en un número finito de puntos, las integrales que se pueden calcular con ella no cambian, por lo tanto, no cambia F_X . Si conocemos F_X y es derivable en un punto x_0 , la convención es tomar $f_X(x_0) = F'_X(x_0)$

Resumen y convención insensata

Si conocemos la distribución $F_X(x)$ obtenemos la densidad derivando:

$$F_X(x) \rightarrow f_X(x) = F'_X(x) .$$

Si conocemos la densidad $f_X(x)$, calculamos la distribución integrando:

$$f_X(x) \rightarrow F_X(x) = \int_{-\infty}^x f_X(u)du .$$

En la literatura del área a las variables aleatorias absolutamente continuas se las denomina, simplemente, continuas. De acá en más, siempre que hablamos de variables aleatorias continuas estaremos queriendo decir variables aleatorias absolutamente continuas.

Definición 4.4 (cuantil α)

Sea X una variable absolutamente continua con función de densidad f_X y función de distribución F_X estrictamente creciente en la región donde $\{0 < F_X < 1\}$. Sea $0 < \alpha < 1$. El α -cuantil (ó cuantil α o α -percentil) de la distribución de X es el valor x_α tal que

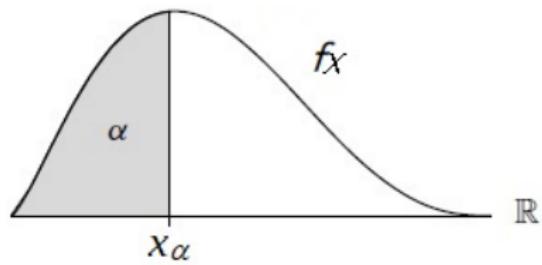
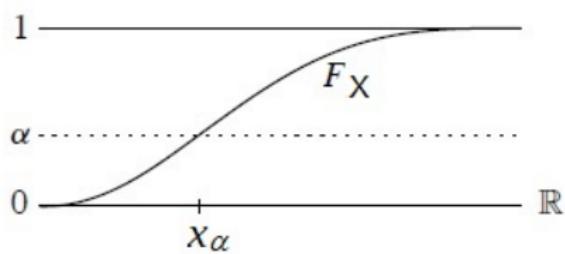
$$F_X(x_\alpha) = P(X \leq x_\alpha) = \int_{-\infty}^{x_\alpha} f_X(u)du = \alpha.$$

Como $\lim_{t \rightarrow -\infty} F_X(t) = 0$, $\lim_{t \rightarrow +\infty} F_X(t) = 1$ y F_X es continua, el α -cuantil de X siempre existe. Como además pedimos que F_X sea estrictamente creciente, es único. Es decir,

$$x_\alpha = F_X^{-1}(\alpha) \Leftrightarrow F_X(x_\alpha) = \alpha.$$

Cuantil α

Figura 1: Cuantil α de una distribución: a la izquierda, graficado en la función de distribución de probabilidad, a la derecha graficado sobre la función de densidad.



Ejercicio 4.1

Sea

$$f_X(x) = \begin{cases} 0 & \text{si } x \leq 0 , \\ \alpha x^2 & \text{si } 0 < x \leq 2 , \\ 0 & \text{si } x > 2 , \end{cases}$$

- ① Hallar α para que f sea una función de densidad. Hallar la función de distribución asociada a la densidad f .
- ② Hallar la probabilidad de que X sea menor a 0.5.
- ③ Hallar el cuantil 0.27 de la distribución.

¿Cuánto difieren "dos densidades"?

¿Cuánto pueden diferir dos "versiones" de la función de densidad de una variable aleatoria?

Definición: Un conjunto $E \subset \mathbb{R}$ tiene medida cero si dado $\epsilon > 0$ existe una familia \mathcal{I} a lo sumo numerable de intervalos

$\mathcal{I} = \{I_j : j \in \mathbb{N}\}$ tal que

$$E \subset \bigcup_{j=1}^{\infty} I_j \quad \text{y} \quad \sum_{j=1}^{\infty} |I_j| < \epsilon$$

donde $|I_j|$ es la longitud del intervalo I_j

Franks, John (2009). A (Terse) Introduction to Lebesgue Integration. The Student Mathematical Library. 48. American Mathematical Society. p. 28. doi:10.1090/stml/048

Propiedades de los conjuntos de medida nula

Propiedades:

- 1) $A \subset B$, B de medida nula $\Rightarrow A$ es de medida nula
- 2) Si $(E_n)_{n \in \mathbb{N}}$ es una sucesión de conjuntos de medida nula, entonces $\bigcup_{n=1}^{\infty} E_n$ también lo es.
- 3) $\{x\}$ es de medida nula. Y por lo tanto, cualquier conjunto numerable es de medida nula.

Caracterización de variables continuas

Proposición: Sea X una r.a. (absolutamente) continua, entonces $P(X \in E) = 0$ para todo $E \in \mathcal{B}(\mathbb{R})$ de medida nula.

(de hecho, vale la equivalencia $\Leftrightarrow X$ es una r.a. continua si y sólo si pero necesitamos resultados de A. Real para probarlo).

$P(X \in E) = 0 \quad \forall E \in \mathcal{B}(\mathbb{R})$ de medida nula.

dem: X es una r.a. continua. Entonces existe una función $f: \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ boreiana (una densidad de X) tal que $P(X \in A) = \int_A f(x) dx \quad \forall A \in \mathcal{B}(\mathbb{R})$.

Sea $\epsilon > 0$. Definimos $C_k = \{x \in \mathbb{R} : f(x) > k\} \in \mathcal{B}(\mathbb{R})$
 $(C_k)_{k \in \mathbb{N}}$ es una sucesión decreciente, $C_{k+1} \subseteq C_k \quad \forall k \in \mathbb{N}$

$$\text{y } \bigcap_{k=1}^{\infty} C_k = \emptyset \quad \therefore \{x \in C_k\} \rightarrow \emptyset$$

$$\text{Luego } P(x \in C_k) \xrightarrow{k \rightarrow \infty} P(\emptyset) = 0$$

Elegimos k_0 tal que $P(x \in C_{k_0}) < \frac{\epsilon}{2}$

Como E tiene medida nula, tomamos $(I_n)_{n \in \mathbb{N}}$ sucesión de intervalos tales que:

$$E \subset \bigcup_{n=1}^{\infty} I_n \quad \text{y} \quad \sum_{n=1}^{\infty} |I_n| < \frac{\epsilon}{2k_0}$$

$$E \subset \bigcup_{n=1}^{\infty} I_n \quad \text{y} \quad \sum_{n=1}^{\infty} |I_n| < \frac{\epsilon}{2k_0}$$

$$\begin{aligned} P(x \in E) &\leq P\left(x \in \bigcup_{n=1}^{\infty} I_n\right) = \underbrace{P\left(x \in \left(\bigcup_{n=1}^{\infty} I_n\right) \cap C_{k_0}\right)} + \underbrace{P\left(x \in \left(\bigcup_{n=1}^{\infty} I_n\right) \cap C_{k_0}^c\right)}_{(*)} \\ &\leq P(x \in C_{k_0}) < \frac{\epsilon}{2} \end{aligned}$$

$$\begin{aligned} (*) &= P\left(x \in \bigcup_{n=1}^{\infty} (I_n \cap C_{k_0}^c)\right) \leq \sum_{n=1}^{\infty} P(x \in I_n \cap C_{k_0}^c) = \sum_{n=1}^{\infty} \int_{I_n \cap C_{k_0}^c} f(x) dx \\ &\leq \sum_{n=1}^{\infty} k_0 \int_{I_n \cap C_{k_0}^c} dx \leq k_0 \sum_{n=1}^{\infty} \int_{I_n} dx = k_0 \underbrace{\sum_{n=1}^{\infty} |I_n|}_{\sum_{n=1}^{\infty} |I_n| < \frac{\epsilon}{2k_0}} < k_0 \frac{\epsilon}{2k_0} = \frac{\epsilon}{2} \end{aligned}$$

¿Cuánto difieren “dos densidades”? Respuesta

Proposición (sin demostración, la verán en Teoría de la medida)

Sea X una v.a. continua y f una densidad de X , f boreiana

Entonces sea $g: \mathbb{R} \rightarrow [0, +\infty)$ boreiana, g es una densidad de X si y sólo si el conjunto $\{x \in \mathbb{R} : f(x) \neq g(x)\}$ tiene medida nula.

De esta proposición se deduce que la densidad f de X es esencialmente única. De ahí el abuso de notación que consiste en denominar a todas estas f posibles de la misma forma, f_X .

Conjuntos de medida cero en \mathbb{R}^k

Para terminar, veamos la definición de conjunto de medida cero en \mathbb{R}^k . Esta definición será útil cuando trabajemos con vectores aleatorios.

Conjunto de medida nula en \mathbb{R}^k

Definición: Un conjunto $E \subset \mathbb{R}^k$ se dice de medida nula si dado $\epsilon > 0$ existe una sucesión $(B_n)_{n \in \mathbb{N}}$ de rectángulos (o productos cartesianos) de lados paralelos a los ejes:

$$B_n = I_1^n \times \dots \times I_k^n \text{ donde } I_i^n \text{ es un intervalo en } \mathbb{R}, 1 \leq i \leq k$$

tal que:

$$E \subset \bigcup_{n=1}^{\infty} B_n \quad \text{y} \quad \sum_{n=1}^{\infty} |B_n| < \epsilon.$$

donde $|B_n|$ es el "volumen" de B_n dado por

$$|B_n| = \prod_{j=1}^k |I_j^n|$$

Variables aleatorias continuas famosas

Les sugiero ver **antes de lo que sigue** los siguientes dos videos:

- sobre variables uniformes

[https:](https://www.youtube.com/watch?v=FAevh_D8Qn0&feature=youtu.be)

[//www.youtube.com/watch?v=FAevh_D8Qn0&feature=youtu.be](https://www.youtube.com/watch?v=FAevh_D8Qn0&feature=youtu.be)

- sobre variables normales

[https:](https://www.youtube.com/watch?v=In7ArRW66NE&feature=youtu.be)

[//www.youtube.com/watch?v=In7ArRW66NE&feature=youtu.be](https://www.youtube.com/watch?v=In7ArRW66NE&feature=youtu.be)

Son para la materia Estadística de químicos, mucho más aplicados que lo que veremos acá, pero transmiten mejor la idea, son más coloquiales (sin abandonar la formalidad). Hablan de esperanza y varianza (que nosotros vamos a tardar unas 3 o 4 semanas en definir, pero pueden saltar esa parte sin perder la idea general). El primer video es sobre la primera transparencia de las que siguen, imagínense lo mejor explicado que está. Son de Mariela Sued.

Variables aleatorias continuas famosas

Uniforme

Diremos que la variable aleatoria X tiene distribución uniforme en el intervalo (a, b) para $a < b$ si su función de densidad esta dada por

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{caso contrario} . \end{cases} \quad (1)$$

Más compacto, $f_X(x) = \frac{1}{b-a} I_{[a,b]}(x)$. Notación: $X \sim \mathcal{U}[a, b]$.

Ejercicio 4.2

Verifique que f_X dada en (1) es una función de densidad.

En tal caso, su función de distribución acumulada de X es

$$F_X(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b , \end{cases}$$

La distribución normal

Definición 4.5

Diremos que la variable aleatoria X tiene distribución normal estándar si su función de densidad es de la forma

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, x \in \mathbb{R}. \quad (2)$$

Notación: $X \sim \mathcal{N}(0, 1)$.

Claramente, f_X definida en (2) es una función positiva. Mostrar que integra uno es un poco más demandante. Requiere integrales múltiples y cambio de variables, como puede verse en el siguiente lema:

Lema 4.4

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1$$

Lema 4.4

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1$$

$$\left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx \right)^2 =$$

Lema 4.4

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1$$

$$\left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx \right)^2 = \left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx \right) \left(\int_{-\infty}^{+\infty} e^{-y^2/2} dy \right)$$

Lema 4.4

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1$$

$$\begin{aligned} \left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx \right)^2 &= \left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx \right) \left(\int_{-\infty}^{+\infty} e^{-y^2/2} dy \right) \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{(x^2+y^2)}{2}} dx dy = \end{aligned}$$

Lema 4.4

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1$$

$$\begin{aligned} \left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx \right)^2 &= \left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx \right) \left(\int_{-\infty}^{+\infty} e^{-y^2/2} dy \right) \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{(x^2+y^2)}{2}} dx dy \end{aligned}$$

Cambio de variables a polares.
 $(x, y) = (r \cos \theta, r \sin \theta)$.

$$0 < r < \infty$$

$$0 < \theta < 2\pi$$

$$\text{Jacobiano} = \left| \det \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \right| = r$$

Lema 4.4

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1$$

$$\begin{aligned} \left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx \right)^2 &= \left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx \right) \left(\int_{-\infty}^{+\infty} e^{-y^2/2} dy \right) \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{(x^2+y^2)}{2}} dx dy = \int_0^{+\infty} \int_0^{2\pi} e^{-\frac{r^2}{2}} r d\theta dr \end{aligned}$$

Cambio de variables a polares.

$$(x, y) = (r \cos \theta, r \sin \theta).$$

$$0 < r < \infty$$

$$0 < \theta < 2\pi$$

$$\text{Jacobiano} = \left| \det \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \right| = r$$

$$= \int_0^{+\infty} \int_0^{2\pi} e^{-r^2/2} r d\theta dr =$$

$$= \int_0^{+\infty} \int_0^{2\pi} e^{-r^2/2} r d\theta dr = 2\pi \int_0^{+\infty} r e^{-r^2/2} dr$$

$$= \int_0^{+\infty} \int_0^{2\pi} e^{-r^2/2} r d\theta dr = 2\pi \int_0^{+\infty} r e^{-r^2/2} dr$$

Cambio var
 \downarrow
 $u = r^2/2$
 $du = r dr$

$$= \int_0^{+\infty} \int_0^{2\pi} e^{-r^2/2} r d\theta dr = 2\pi \int_0^{+\infty} r e^{-r^2/2} dr$$

Cambio var
 \downarrow
 $u = r^2/2$
 $du = r dr$

$$= 2\pi \int_0^{+\infty} e^{-u} du =$$

$$= \int_0^{+\infty} \int_0^{2\pi} e^{-r^2/2} r d\theta dr = 2\pi \int_0^{+\infty} r e^{-r^2/2} dr$$

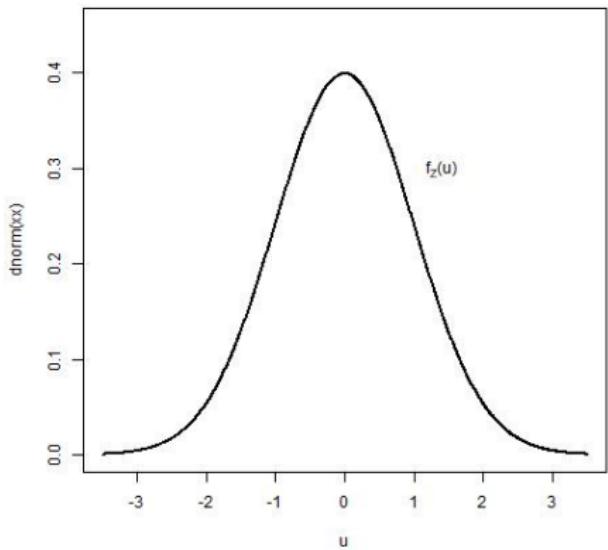
Cambio var
 \downarrow
 $u = r^2/2$
 $du = r dr$

$$= 2\pi \int_0^{+\infty} e^{-u} du = 2\pi (e^{-u}) \Big|_0^{+\infty} = 2\pi \quad \text{⊗}$$

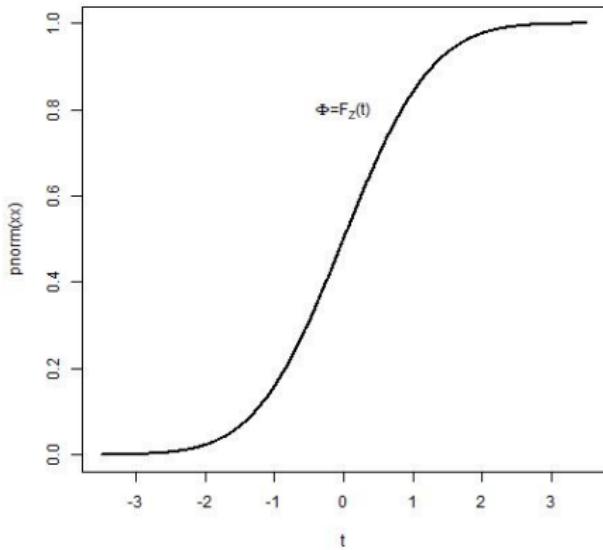
No es posible obtener una fórmula explícita para la función de distribución acumulada de una variable aleatoria normal estándar. Métodos numéricos han permitido construir una tabla con los valores de la función de distribución acumulada para diferentes valores de x . En la bibliografía, la función de distribución acumulada de una variable aleatoria normal estándar es denotada por

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du .$$

Densidad y distribución normal estándar



Función de densidad de la normal estándar, o campana de Gauss.



Función de distribución acumulada (Φ), para una v.a. $Z \sim \mathcal{N}(0, 1)$

Ejercicio 4.3

Sea $Z \sim \mathcal{N}(0, 1)$. Consideremos $X = \sigma Z + \mu$, con $\sigma > 0$. ¿Cuál es la función de densidad de la variable aleatoria X ?

(Sugerencia: Expresar F_X en función de F_Z y luego derivar)

$$F_X(x) = P(X \leq x) = P(\sigma Z + \mu \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = F_Z\left(\frac{x-\mu}{\sigma}\right)$$

Derivando tenemos

$$F'_X(x) = f_X(x) = f_Z\left(\frac{x-\mu}{\sigma}\right) \frac{1}{\sigma} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, x \in \mathbb{R}$$

Definición 4.6 ($N(\mu, \sigma^2)$)

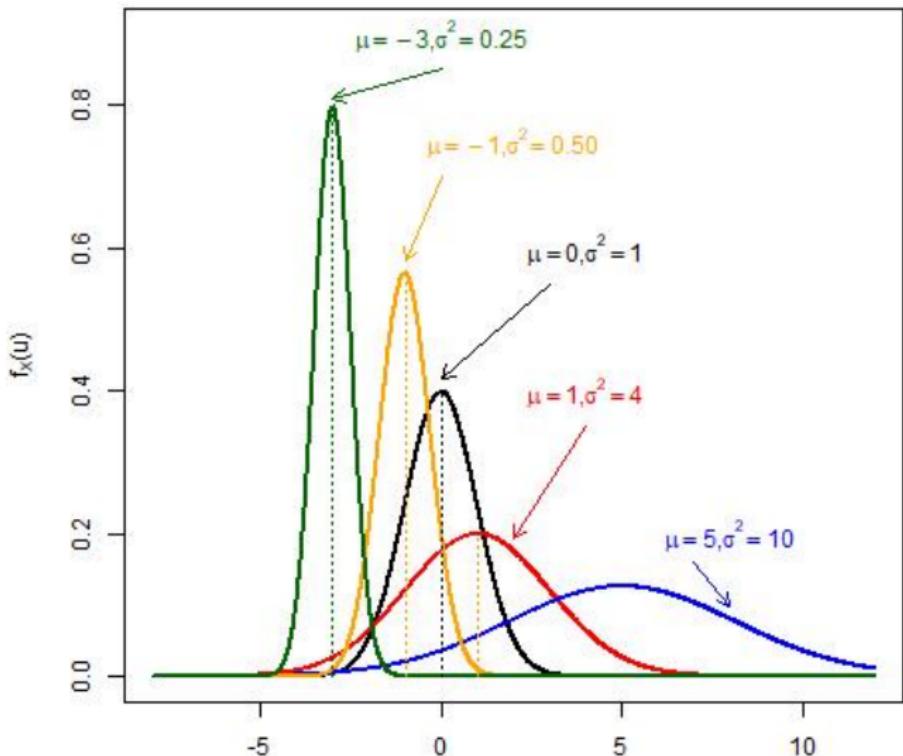
Diremos que la variable aleatoria X tiene **distribución normal de parámetros μ y σ^2** si su función de densidad es de la forma

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(x-\mu)^2/2\sigma^2\right\}, x \in \mathbb{R}. \quad (3)$$

Notación: $X \sim N(\mu, \sigma^2)$.

Obs: Cuando $\mu = 0$ y $\sigma^2 = 1$ tenemos una $\mathcal{N}(0, 1)$ ó normal estándar.

Función de densidad $N(\mu, \sigma^2)$ para varios μ y σ^2



Lema 4.5

Sea $X \sim \mathcal{N}(\mu, \sigma^2)$. Definimos $Y = \alpha X + \beta$, con $\alpha \neq 0$. Entonces, la distribución de Y es normal con parámetros $\alpha\mu + \beta$ y $\alpha^2\sigma^2$:
 $Y \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$.

Demostración.

Para demostrar este hecho, supongamos $\alpha > 0$ y calculemos la función de distribución de la variable aleatoria Y :

$$F_Y(y) = P(Y \leq y) = P(\alpha X + \beta \leq y) = P\left(X \leq \frac{y - \beta}{\alpha}\right) = F_X\left(\frac{y - \beta}{\alpha}\right).$$

Derivando con respecto a y ambos miembros, obtenemos que la función de densidad de Y corresponde a una normal con los parámetros anunciados (comprobarlo).

Ejercicio el caso $\alpha < 0$.



Corolario 4.6 (estandarización de la normal)

En particular, tenemos que si $X \sim \mathcal{N}(\mu, \sigma^2)$, entonces

$Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$ tiene distribución normal estándar. Recordando que $\Phi(\cdot)$ denota la función acumulada de una variable Z normal estándar cuyos valores están tabulados, tenemos que

- Para calcular probabilidades que involucran a X tenemos

$$\begin{aligned} F_X(x) &= P(X \leq x) = P\left(\frac{(X - \mu)}{\sigma} \leq \frac{(x - \mu)}{\sigma}\right) \\ &= P\left(Z \leq \frac{(x - \mu)}{\sigma}\right) = \Phi\left(\frac{(x - \mu)}{\sigma}\right) \end{aligned}$$

Corolario 4.6

- Para calcular cuantiles de X , $x_\alpha = F_X^{-1}(\alpha)$ y $z_\alpha = \Phi^{-1}(\alpha)$, pues Φ es estrictamente creciente y continua. ¿Cómo se relacionan?

$$\begin{aligned}\alpha &= F_X(x_\alpha) = P(X \leq x_\alpha) = P\left(\frac{(X - \mu)}{\sigma} \leq \frac{(x_\alpha - \mu)}{\sigma}\right) \\ &= P\left(Z \leq \frac{(x_\alpha - \mu)}{\sigma}\right) = \Phi\left(\frac{(x_\alpha - \mu)}{\sigma}\right).\end{aligned}$$

Luego

$$\frac{(x_\alpha - \mu)}{\sigma} = \Phi^{-1}(\alpha) = z_\alpha$$

que está tabulado. O equivalentemente,

$$x_\alpha = z_\alpha \sigma + \mu$$

En R: Normal

En R, la función `dnorm` calcula la función de densidad de la distribución normal:

`dnorm(x, mean, sd)` calcula la $f_X(x)$ para $X \sim \mathcal{N}(\text{mean}, \text{sd}^2)$.

`pnorm(x, mean, sd)` calcula la $F_X(x)$ para dicha distribución.

`qnorm(p, mean, sd)` calcula el p -cuantil de la distribución de $X \sim \mathcal{N}(\text{mean}, \text{sd}^2)$. Si no se le dan los valores del segundo y tercer argumento, tanto para la densidad, distribución o cuantil, devuelve los valores correspondientes a la normal estándar. Por ejemplo,

```
> pnorm(0)
```

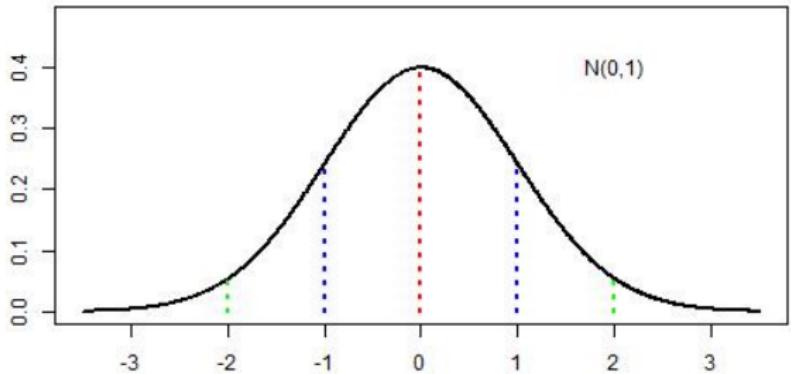
```
[1] 0.5
```

```
> qnorm(0.5)
```

```
[1] 0
```

De hecho, `pnorm` y `qnorm` son funciones inversas.

Estandarización: gráficos



El gráfico superior corresponde a la densidad de una $Z \sim N(0, 1)$, el gráfico inferior vemos la densidad de una $X \sim N(\mu, \sigma^2)$.

```
> pnorm(1)-pnorm(-1)
```

```
[1] 0.6826895
```

```
> pnorm(2)-pnorm(-2)
```

```
[1] 0.9544997
```

```
> pnorm(3)-pnorm(-3)
```

```
[1] 0.9973002
```

En el primer gráfico vemos que, por ejemplo,
 $P(-1 < Z < 1) = 0,683$
que es igual a la
 $P(\mu - \sigma < X < \mu + \sigma)$ en el segundo.

Distribución exponencial

Definición 4.7 (Distribución exponencial)

Diremos que X tiene distribución exponencial de parámetro $\lambda > 0$ si su función de densidad está dada por

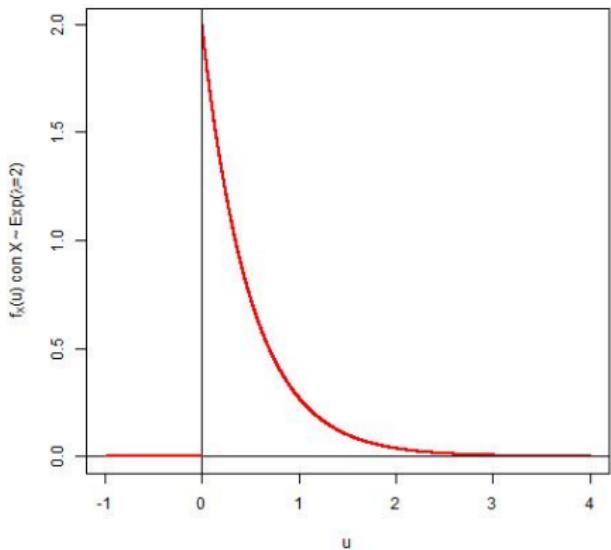
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{caso contrario.} \end{cases}$$

Es decir, $f_X(x) = \lambda e^{-\lambda x} I_{[0,+\infty)}(x)$. En tal caso, $F_X(x) = 0$ para $x < 0$ y $F_X(x) = 1 - e^{-\lambda x}$ para $x \geq 0$. Notación: $X \sim \mathcal{E}(\lambda)$.

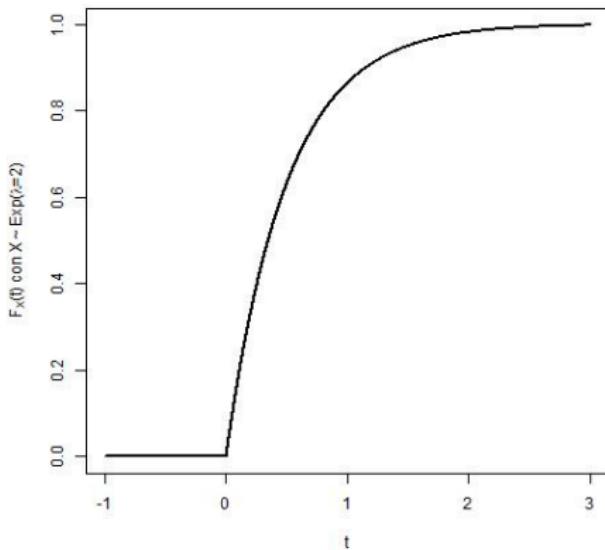
Las variables aleatorias exponenciales suelen utilizarse para modelar tiempo de vida de artefactos con falta de memoria, debido a la siguiente propiedad que, salvo caso trivial, sólo la distribución exponencial verifica.

Densidad y distribución exponencial

$\mathcal{E}(\lambda)$ con $\lambda = 2$.



Función de densidad.



Función de distribución acumulada.

Proposición 4.1 (Propiedad de falta de memoria)

Sea $X \sim \mathcal{E}(\lambda)$. Entonces,

$$P(X > t + s | X > t) = P(X > s) \text{ para todo } s, t > 0 \quad (4)$$

Recíprocamente, si X es una variable aleatoria no negativa satisfaciendo la propiedad (4), entonces X es idénticamente nula ($P(X = 0) = 1$) ó $X \sim \mathcal{E}(\lambda)$, para algún $\lambda > 0$.

Demostración.

Para verificar que la distribución exponencial verifica lo pedido, basta notar que si $X \sim \mathcal{E}(\lambda)$, entonces $P(X > t) = e^{-\lambda t}$, en cuyo caso $P(X > t + s) = P(X > t)P(X > s)$.

Sea ahora X una variable aleatoria no negativa tal que verifique la propiedad de falta de memoria. En tal caso, si ponemos $P(X > t) = G(t)$, tenemos que $G(t) = 1$ para todo $t < 0$, G es continua a derecha y verifica

$$G(t + s) = G(t)G(s), \forall s, t > 0. \quad (5)$$

De la identidad (5) podemos probar que, $\forall n, m \in \mathbb{N}$,

1) $G(n) = \{G(1)\}^n$. Dem: $G(n) = G(\underbrace{1 + 1 + \cdots + 1}_{n \text{ sumandos}}) = \{G(1)\}^n$

2) $G(\frac{1}{m}) = \{G(1)\}^{\frac{1}{m}}$. Dem: $G(1) = G\left(\underbrace{\frac{1}{m} + \cdots + \frac{1}{m}}_{m \text{ sumandos}}\right) = G\left(\frac{1}{m}\right)^m$



Demostración.

3. $G(q) = \{G(1)\}^q$ para todo $q \in \mathbb{Q}_{>0}$. Dem: Como $q = \frac{n}{m}$, luego
 $G\left(\frac{n}{m}\right) = G\left(\underbrace{\frac{1}{m} + \cdots + \frac{1}{m}}_{n \text{ sumandos}}\right) = [G\left(\frac{1}{m}\right)]^n = [G(1)]^{\frac{n}{m}} = \{G(1)\}^q$.

Como G continua a derecha, concluimos que $G(t) = \{G(1)\}^t$ para todo $t \in \mathbb{R}_{\geq 0}$.

Ahora bien, como $G(1) = P(X > 1)$, tenemos que $0 \leq G(1) \leq 1$.

Si $G(1) = 1$, $G(t) = 1$ para todo $t \geq 0$, con lo cual

$F_X(t) = P(X \leq t) = 0$ para todo $t \geq 0$, cosa que no puede ocurrir.

Por otra parte, si $G(1) = 0$ tenemos que $F_X(t) = 1$ para todo $t \geq 0$. Como además sabemos que X es no negativa, concluimos que $P(X = 0) = 1$.

Finalmente, si $0 < G(1) < 1$, poniendo $\lambda = -\ln(G(1))$, tenemos que $\lambda > 0$ y

$$G(t) = P(X > t) = e^{-\lambda t}$$

para $t \geq 0$ y $G(t) = 1$ para $t < 0$, lo cual prueba que $X \sim \mathcal{E}(\lambda)$.



Falta de memoria

$$P(X > t + s \mid X > t) = P(X > s) \text{ para todo } s, t > 0$$

La falta de memoria dice que si X es la variable aleatoria que mide la duración de un objeto (la vida útil de la batería de un auto), sabiendo que su edad es t días, la probabilidad de que “dure” más de s días es la misma que la probabilidad inicial de que dure s días. No sería adecuada la distribución exponencial para modelar la va

X = duración de la vida de una persona,
ya que, por ejemplo sabemos que

$$P(X > 90 + 20 \mid X > 90) \neq P(X > 20)$$

Distribución Gama

Recordemos, de Análisis I, la **función Gama**, $\Gamma : \mathbb{R}_{>0} \rightarrow \mathbb{R}$ dada por

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx. \quad (6)$$

Esta función definida mediante la fórmula precedente está bien definida (como integral impropia) y goza de las siguientes propiedades:

Propiedades

- 1) Integrando por partes, obtenemos que $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.
- 2) Siendo $\Gamma(1) = 1$, de la propiedad anterior tenemos que $\Gamma(n) = (n - 1)!$ para todo $n \in \mathbb{N}$. Es por ello que decimos que la función Γ extiende el concepto de factorial para números positivos.
- 3) Haciendo el cambio de variables $x = \frac{u^2}{2}$, puede mostrarse que $\Gamma(1/2) = \sqrt{\pi}$.

Distribución Gama

Definición 4.8 (Distribución Gama)

Diremos que la variable aleatoria X tiene **distribución Gama** con parámetros $\alpha > 0$ y $\lambda > 0$ si su función de densidad viene dada por la fórmula

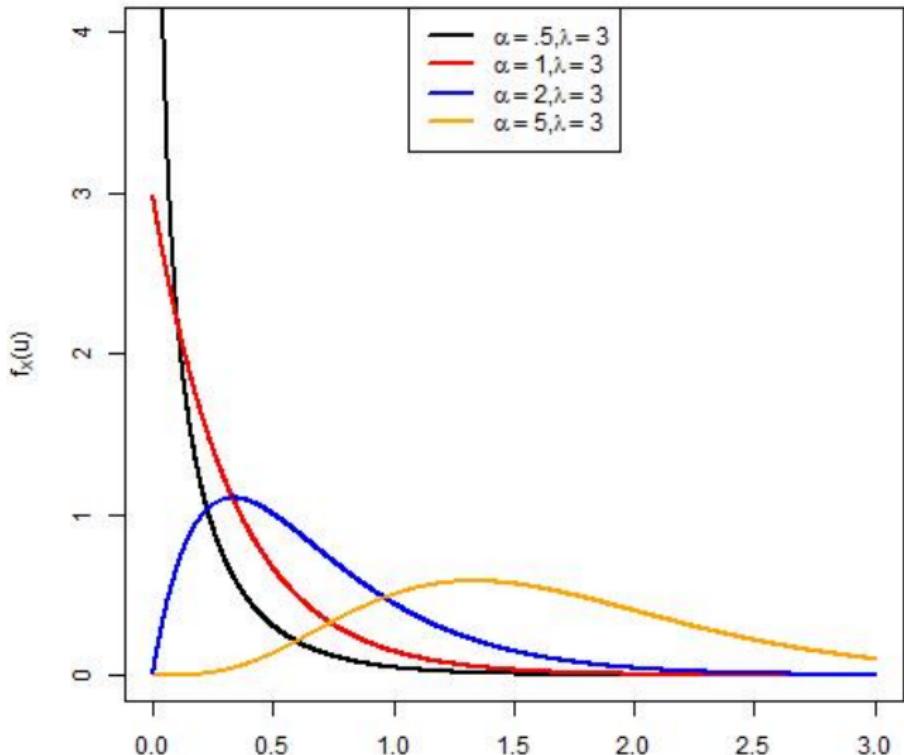
$$f_X(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{caso contrario.} \end{cases} \quad (7)$$

Notación: $X \sim \Gamma(\alpha, \lambda)$.

Observación 4.4

- Si $\alpha = 1$, $\Gamma(1, \lambda) = \mathcal{E}(\lambda)$.
- Cuando $\alpha = \frac{n}{2}$, para algún n entero y $\lambda = \frac{1}{2}$, a la distribución $\Gamma(\frac{n}{2}, \frac{1}{2})$ se la denomina **Chi-cuadrado con n grados de libertad**, $\chi^2(n)$. Esta distribución juega un papel importante en estadística.

Función de densidad $\Gamma(\alpha, \lambda)$ para varios α y $\lambda = 3$



Distribución Beta

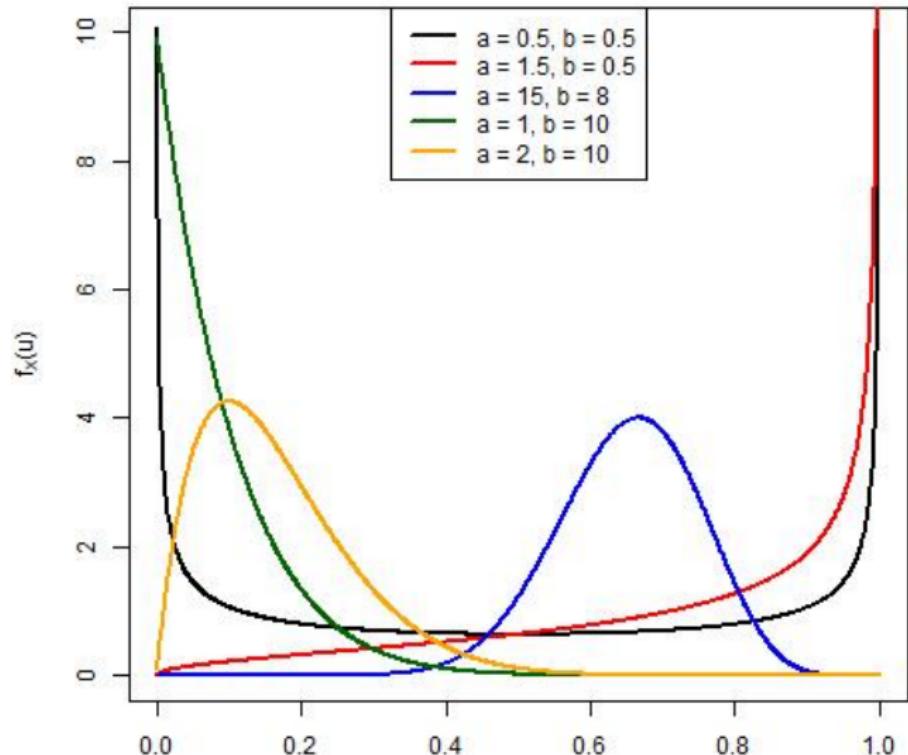
Definición 4.9

Se define la *distribución beta con parámetros a y b* $a, b > 0$ que denotaremos por $\beta(a, b)$, como la distribución absolutamente continua cuya función de densidad es:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} I_{(0,1)}(x).$$

La constante $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ asegura que la integral de f es igual a 1. Más adelante en la materia, en la práctica van a hacer una cuenta que les permitirá comprobarlo. Si $a = b = 1$, tenemos $\beta(a, b) = \mathcal{U}(0, 1)$.

Función de densidad $\beta(a, b)$, para varios a y b



Funciones de variables aleatorias

Queremos construir nuevas variables aleatorias, a partir de otras. Sea X una variable aleatoria, nos interesa $Y = g(X)$ con $g : \mathbb{R} \rightarrow \mathbb{R}$. Ya tratamos un caso simple de esto cuando vimos

$$Y = aX + b, \quad \text{con } X \text{ una variable aleatoria normal}$$

¿Cuándo será $Y = g(X)$ una nueva variable aleatoria? Necesitamos que

$$Y^{-1}((-\infty, a]) = \{Y \leq a\} \in \mathcal{F} \quad \forall a \in \mathbb{R}.$$

Como $Y = g \circ X$, la preimagen cumple

$$Y^{-1}((-\infty, a]) = X^{-1}(g^{-1}((-\infty, a]))$$

Basta que $g^{-1}((-\infty, a]) \in \mathcal{B}(\mathbb{R}) \quad \forall a \in \mathbb{R}$ para tener garantizado que Y es una nueva variable aleatoria.

Funciones Borelianasy

Definición 4.10 (Funciones Borelianasy)

Diremos que una función $g : A \subset \mathbb{R} \rightarrow \mathbb{R}$ es *boreiana* (o medible Borel) si $g^{-1}((-\infty, a]) \in \mathcal{B}(\mathbb{R}) \quad \forall a \in \mathbb{R}$

Proposición 4.2

$g : \mathbb{R} \rightarrow \mathbb{R}$ es boreiana si y sólo si $g^{-1}(B) \in \mathcal{B}(\mathbb{R}) \quad \forall B \in \mathcal{B}(\mathbb{R})$

Demostración.

\Leftrightarrow) Trivial pues $(-\infty, a] \in \mathcal{B}(\mathbb{R}) \quad \forall a \in \mathbb{R}$.

\Rightarrow) Es análoga a cuando probamos la propiedad análoga para variables aleatorias. La familia de conjuntos $\mathcal{G} = \{B \in \mathcal{B}(\mathbb{R}) : g^{-1}(B) \in \mathcal{F}\}$ satisface

- ① es una σ -álgebra en \mathbb{R}
- ② Contiene a las semirrectas de la forma $\{(-\infty, a], a \in \mathbb{R}\} = \mathcal{I}$.
- ③ Por lo tanto, contiene a la σ -álgebra generada por ellas, $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{I}) \subset \mathcal{G}$. Y listo.

Ejemplos de funciones boreelianas

① **Funciones continuas** Dem: La preimagen por una función continua de un conjunto $F = (-\infty, a] \subset \mathbb{R}$ cerrado es cerrado, y por lo tanto boreiano.

② **Funciones monótonas crecientes**

Dem: Sea $x_a = \sup \{x \in \mathbb{R} : g(x) \leq a\}$. Entonces:

- (i) si $g(x_a) \leq a$, $g^{-1}((-\infty, a]) = (-\infty, x_a] \in \mathcal{B}(\mathbb{R})$,
- (ii) si $g(x_a) > a$, $g^{-1}((-\infty, a]) = (-\infty, x_a) \in \mathcal{B}(\mathbb{R})$.

③ **Funciones monótonas decrecientes** Dem: análoga.

④ **Función indicadora de un boreiano** Dem: si $g = I_A$, entonces $g^{-1}(B) \in \{\emptyset, A, A^c, \mathbb{R}\} \subset \mathcal{B}(\mathbb{R})$.

⑤ **Dadas una sucesión de funciones boreelianas, $\{g_n\}_{n \in \mathbb{N}}$ entonces también son boreianas:**

$$\sup_n g_n, \inf_n g_n, \limsup_{n \rightarrow \infty} g_n, \liminf_{n \rightarrow \infty} g_n.$$

- 5) Dadas una sucesión de funciones borelianasy, $\{g_n\}_{n \in \mathbb{N}}$ entonces también son boreianas: $\sup_n g_n$, $\inf_n g_n$, $\limsup_n g_n$, $\liminf_n g_n$.

Demostración.

Como el supremo de una sucesión es $\leq a$ si y sólo si todos los elementos de la sucesión son $\leq a$ tenemos

$$\left\{ \sup_n g_n \leq a \right\} = \bigcap_n \{g_n \leq a\} \in \mathcal{B}(\mathbb{R})$$

porque cada $\{g_n \leq a\} \in \mathcal{B}(\mathbb{R})$. Un argumento similar prueba $\{\inf_n g_n < a\} = \bigcup_n \{g_n < a\} \in \mathcal{B}(\mathbb{R})$, y el resultado se deduce de que la σ -álgebra $\mathcal{B}(\mathbb{R})$ también está generada por las semirrectas de la forma $(-\infty, a)$. Para las otras dos, recordemos

$$\begin{aligned}\liminf_{n \rightarrow \infty} g_n &= \sup_n (\inf_{m \geq n} g_m) \\ \limsup_{n \rightarrow \infty} g_n &= \inf_n (\sup_{m \geq n} g_m)\end{aligned}$$



Entonces, si X es una variable aleatoria, X^2 , e^X , $|X|$, $g(X)$ también lo son, para toda g continua.

Nos interesa ahora decir algo de la distribución de $Y = g(X)$ si conocemos la distribución de X .

Ejemplo 4.7

Sea X una variable (absolutamente) continua con función de densidad f_X .

Sea $g : \mathbb{R} \rightarrow \mathbb{R}$ definida por $g(x) = e^x$. Hallar la función de densidad de $Y = g(X) = e^X$.

Sugerencia: Expresar a F_Y usando F_X y después derivar.

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) \underset{\text{si } y > 0}{=} P(X \leq \ln(y)) = F_X(\ln(y)).$$

Observar que si $y \leq 0$, $P(e^X \leq y) = 0$. Luego,

$$F_Y(y) = F_X(\ln(y))I_{(0,+\infty)}(y).$$

Derivando con respecto a y tenemos:

$$f_Y(y) = f_X(\ln(y)) \frac{1}{y} I_{(0,+\infty)}(y).$$

Teorema 4.8

Sea X una v.a. con densidad $f_X(x)$ tal que $P(X \in (a, b)) = 1$. Sea $g : (a, b) \rightarrow \mathbb{R}$ **estrictamente creciente o bien estrictamente decreciente**, derivable y con $g'(y) \neq 0$. Sea $Y = g(X)$. Entonces Y es una variable aleatoria continua con función de densidad

$$f_Y(y) = f_X(g^{-1}(y)) |(g^{-1})'(y)|.$$

Demostración.

(caso creciente)

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

y derivamos

$$f_Y(y) = f_X(g^{-1}(y)) (g^{-1})'(y).$$

(caso decreciente) Ejercicio.



Funciones no inyectivas de variables aleatorias

Muchas veces, pese a que la función g no es inversible, podemos calcular la función de densidad de $Y = g(X)$. Por ejemplo,

Ejercicio 4.4

- Sea X una variable continua con densidad f_X y sea $Y = X^2$. Calcule $F_Y(y)$, la función de distribución acumulada de Y y $f_Y(y)$, la densidad de Y .
- Consideremos $X \sim \mathcal{U}[-3, 3]$ y sea $Y = X^2$. Calcule $f_Y(y)$.

Ejercicio 4.5

Sea $Z \sim \mathcal{N}(0, 1)$ y sea $Y = Z^2$. Calcule $f_Y(y)$, la densidad de Y . ¿A qué familia pertenece? ¿Con qué parámetros?

Inversa generalizada

Entre las funciones $g(X)$ de una variable aleatoria X hay una que nos interesa por sus propiedades probabilísticas. Estamos interesados en la variable aleatoria $Y = F_X^{-1}(X)$. O sea: X es una variable aleatoria.

Conocemos su función de distribución F_X calculamos su inversa " F_X^{-1} " (que no siempre está bien definida pero proponemos una versión que funcione más o menos como la inversa). Entonces, " F_X^{-1} " : $(0, 1) \rightarrow \mathbb{R}$

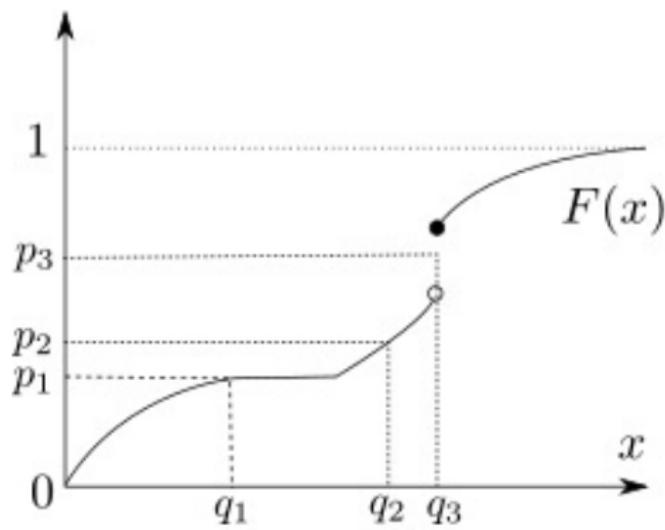
luego definimos la variable aleatoria

$$Y = "F_X^{-1}"(X)$$

y veremos

un teorema que prueba un resultado importante. Ya nos cruzamos con la función inversa cuando definimos los cuantiles de una distribución absolutamente continua. Primero nos concentraremos en definir la inversa generalizada de una función de distribución $F : \mathbb{R} \rightarrow (0, 1)$, que en el caso invisible coincide con la inversa.

Inversa generalizada



Hay **dos problemas** para definir la $F^{-1}(y)$ con $y \in (0, 1)$ para F una función de distribución que ilustramos en el gráfico.

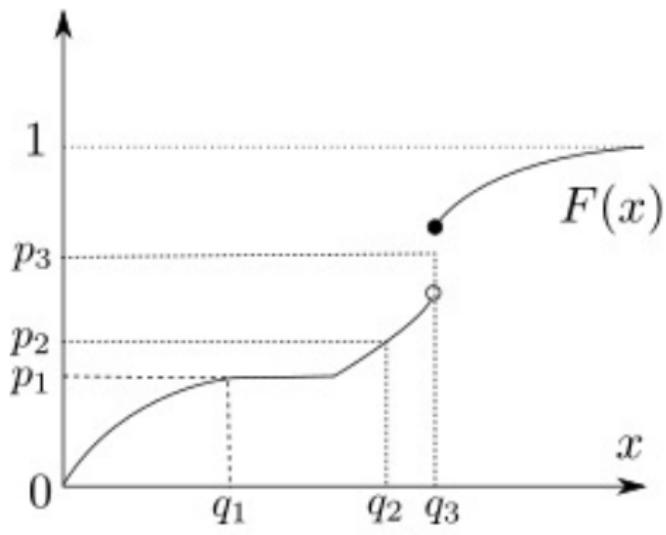
- ① $p_3 \notin \text{Imagen de } F$. O sea $F^{-1}(\{p_3\}) = \emptyset$. (F es discontinua)
- ② En cambio, $F^{-1}(\{p_1\})$ contiene muchos valores y habrá que elegir uno.

Finalmente, como sea que definamos la inversa generalizada debe verificar que $F^{-1}(p_2) = q_2$

Inversa generalizada

Para $0 < y < 1$ sea

$$A_y = \{x \in \mathbb{R} : F(x) \geq y\}$$



En el gráfico

- ① $A_{p_3} = [q_3, +\infty)$
- ② $A_{p_2} = [q_2, +\infty)$
- ③ $A_{p_1} = [q_1, +\infty)$

Candidato a ser una buena inversa es definir:

$$F^{-1}(y) = \inf A_y$$

Lema 4.9

Dada F una función de distribución, tenemos que

- ① $A_y \neq \emptyset$, siendo que $F(x) \rightarrow 1$ cuando $x \rightarrow \infty$.
- ② A_y acotado inferiormente, siendo F creciente y $F(x) \rightarrow 0$ cuando $x \rightarrow -\infty$.

Definición 4.11 (Inversa generalizada)

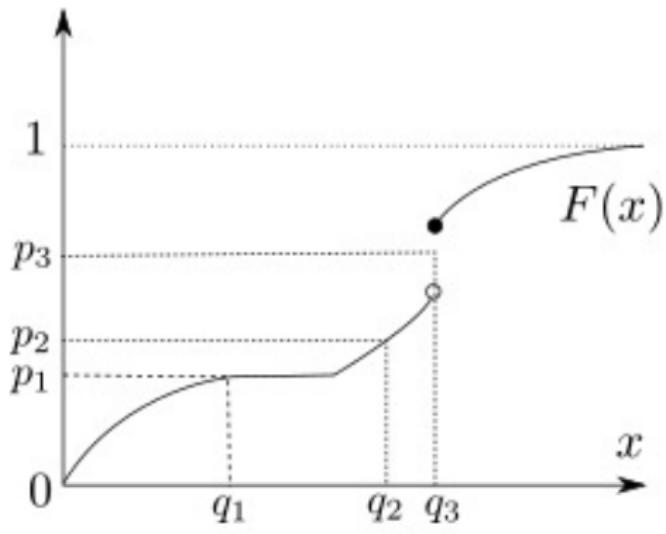
Dada una función de distribución F , denotaremos por F^{-1} a la función que para $y \in (0, 1)$ toma el valor

$$F^{-1}(y) = \inf A_y. \quad (8)$$

Inversa generalizada

Para $0 < y < 1$ sea

$$A_y = \{x \in \mathbb{R} : F(x) \geq y\}$$



$$F^{-1}(y) = \inf A_y$$

En el gráfico

- ① $A_{p_3} = [q_3, +\infty), F^{-1}(p_3) = q_3$
- ② $A_{p_2} = [q_2, +\infty), F^{-1}(p_2) = q_2$
- ③ $A_{p_1} = [q_1, +\infty), F^{-1}(p_1) = q_1$

Lema 4.10 (Propiedades de la inversa generalizada)

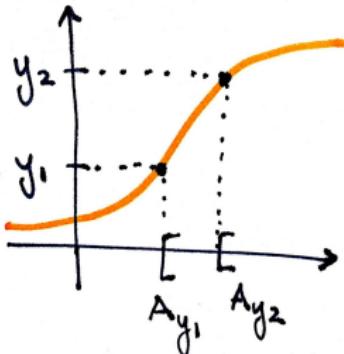
- ① El ínfimo de A_y es en realidad un mínimo: $F^{-1}(y) \in A_y$, $0 < y < 1$.
- ② F^{-1} creciente (en sentido amplio).
- ③ Si F es continua, $F(F^{-1}(y)) = y$
- ④ $F^{-1}(F(x)) \leq x$. Es decir, $x \in A_{F(x)}$.
- ⑤ $F^{-1}(y) \leq x \Leftrightarrow y \leq F(x)$

Demostración.

- ① Sea $x_n \in A_y$ decreciendo al ínfimo: $x_n \searrow F^{-1}(y)$. Siendo F continua a derecha, tenemos que $F(x_n) \rightarrow F(F^{-1}(y))$. Pero como $x_n \in A_y$, $F(x_n) \geq y$ para todo n , y entonces lo mismo ocurre con $F(F^{-1}(y))$. Observemos que este ítem es equivalente a que

$$F(F^{-1}(y)) \geq y .$$





Demostración.

2) [Qvq F^{-1} creciente] Sea $y_1 < y_2$. Luego $A_{y_2} \subset A_{y_1}$ (comprobarlo) (ver gráfico). Por propiedades de ínfimo, $F^{-1}(y_1) \leq F^{-1}(y_2)$.

□

Demostración.

3) [Qvq Si F es continua, $F(F^{-1}(y)) = y$] Por 1) tenemos $y \leq F(F^{-1}(y))$. Si fuera $y < F(F^{-1}(y))$, existiría y^* con $y < y^* < F(F^{-1}(y))$. Como F es continua, por Bolzano, existe x^* tal que $y^* = F(x^*)$, es decir

$$y < F(x^*) < F(F^{-1}(y)). \quad (9)$$

Luego $x^* \in A_y$, y por propiedades de ínfimo, $F^{-1}(y) \leq x^*$. Como F es creciente, si la aplicamos a la última desigualdad tenemos $F(F^{-1}(y)) \leq F(x^*)$, lo cual contradice (9). Luego, $y = F(F^{-1}(y))$

□

Demostración.

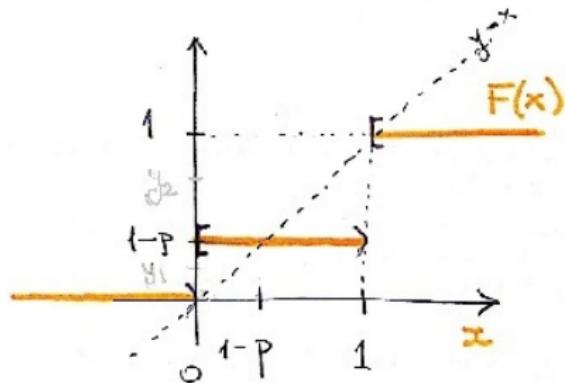
4) [Qvq $F^{-1}(F(x)) \leq x$. Es decir, $x \in A_{F(x)}$.] Por definición de $A_{F(x)}$ tenemos que $x \in A_{F(x)}$. El resultado sigue de que $F^{-1}(F(x)) = \inf A_{F(x)}$.

5) [Qvq $F^{-1}(y) \leq x \Leftrightarrow y \leq F(x)$.]
⇒) $F^{-1}(y) \leq x$. Como F es creciente resulta $F(F^{-1}(y)) \leq F(x)$.
Pero, de 1) tenemos que $y \leq F(F^{-1}(y))$.
⇐) $y \leq F(x)$. Luego, como por 2) F^{-1} es creciente entonces
 $F^{-1}(y) \leq F^{-1}(F(x))$ $\underbrace{\leq}_{\text{por 4)}} x$



Ejercicio 4.6

Hallar la inversa generalizada de la función de distribución acumulada de una variable aleatoria $Be(p)$. Graficar en el mismo eje de coordenadas a F y a F^{-1} , pero recordando que el dominio de la segunda es el $(0,1)$.



$$F : \mathbb{R} \rightarrow [0,1]$$

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1-p & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

$$A_{y_2} = \{x \in \mathbb{R} : F(x) \geq y_2\} = \{x : F(x) \geq 1\} = [1, +\infty) = A_1$$

$\forall y_2 \in (1-p, 1)$

$$A_{y_1} = \{x \in \mathbb{R} : F(x) \geq y_1\} = \{x : F(x) \geq 1-p\} = [0, +\infty) = A_0$$

$\forall y_1 \in (0, 1-p]$

$$F^{-1} : (0,1) \rightarrow \mathbb{R}$$

$$F^{-1}(y) = \begin{cases} 0 & \text{si } 0 < y \leq 1-p \\ 1 & \text{si } 1-p < y < 1 \end{cases}$$

Proposición 4.3 (Generación de números al azar)

Sea $U \sim \mathcal{U}(0, 1)$, y F una función de distribución. Entonces, $X = F^{-1}(U)$ es una variable aleatoria con función de distribución F .

Demostración.

F una función de distribución, i.e. satisface las propiedades del Lema 3.3:

- $F : \mathbb{R} \rightarrow [0, 1]$
- es monótona creciente (en sentido amplio),
- continua a derecha con límite a izquierda
- $\lim_{t \rightarrow -\infty} F_X(t) = 0$ y $\lim_{t \rightarrow +\infty} F_X(t) = 1$.

Entonces está bien definida F^{-1} la inversa generalizada, que resulta creciente (en sentido amplio) y por lo tanto boreiana. Entonces $X = F^{-1}(U)$ es una variable aleatoria. Luego

$$F_X(x) = P(X \leq x) = P(F^{-1}(U) \leq x) \quad \underbrace{=} \quad P(U \leq F(x)) = F(x).$$

por Lema 4.10 5)

Generación de números al azar

La proposición que acabamos de probar es también la demostración constructiva del Lema 3.4 que habíamos enunciado la semana pasada:

Lema 3.4

Si $F : \mathbb{R} \rightarrow [0, 1]$ y satisface las propiedades 2), 3) y 4) del Lema 3.3 entonces existe una variable aleatoria X definida en un espacio de probabilidad (Ω, \mathcal{F}, P) tal que $F_X = F$.

En este caso, la X que definimos está definida en el mismo espacio de probabilidad que la uniforme, $U : (\Omega, \mathcal{F}, P) \rightarrow \mathbb{R}$ entonces $X : (\Omega, \mathcal{F}, P) \rightarrow \mathbb{R}$.

Luego, la proposición nos permite construir una variable aleatoria X con la distribución que queramos si sabemos construir una v.a. con distribución $\mathcal{U}(0, 1)$. Es lo mismo que saber elegir un número al azar en el intervalo $(0, 1)$. ¿Cómo se genera una realización de una v.a. $\mathcal{U}(0, 1)$?

¿Cómo generar realizaciones de una v.a. $\mathcal{U}(0, 1)$?

H.-O. Georgii. *Stochastics: introduction to probability and statistics*. Walter de Gruyter, 2012, pág 81.

(3.45) Remark. *Random numbers.* Random realisations of independent $\mathcal{U}_{[0,1]}$ -distributed random variables are called *random numbers*. They can be found in tables or on the internet. Partly, these are generated by real chance, see e.g. http://www.rand.org/pubs/monograph_reports/MR1418/index.html. In practice, however, it is common to use so-called *pseudo-random numbers*, which are not random at all, but produced deterministically. A standard method to generate pseudo-random numbers is the following *linear congruence method*.

Choose a ‘modulus’ m (for example $m = 2^{32}$) as well as a factor $a \in \mathbb{N}$ and an increment $b \in \mathbb{N}$ (this requires a lot of skill). Next, pick a ‘seed’ $k_0 \in \{0, \dots, m - 1\}$ (e.g., tied to the internal clock of the processor), and define the recursive sequence $k_{i+1} = ak_i + b \bmod m$. The pseudo-random numbers then consist of the sequence $u_i = k_i/m$, $i \geq 1$. For an appropriate choice of a, b , the sequence (k_i) has period m (so it does not repeat itself after fewer iterations), and it survives several statistical tests of independence. (For example, this is the case for $a = 69069$, $b = 1$; Marsaglia 1972.)

Referencia standard sobre números seudo aleatorios: Knuth, D.E. *The art of computer programming*, Vol2, 1997.

Generar realizaciones de una v.a. $\mathcal{U}(0, 1)$ en R

Generamos un vector que llamaremos **uu** de longitud 1000, cada coordenada de **uu**, que llamaremos U_i , es el resultado de elegir un número al azar en el intervalo $(0, 1)$ en R. Es decir, cada coordenada es la realización de una v.a. con distribución $\mathcal{U}(0, 1)$. Después mostramos los 10 primeros obtenidos.

```
#-----
#     Generacion de numeros al azar
#-----
> set.seed(07052020)
> uu<-runif(1000)
> uu[1:10]
[1] 0.8244874 0.3245704 0.1708590 0.9438833 0.3122139
[6] 0.5936935 0.1067565 0.7206148 0.9457670 0.7539267
```

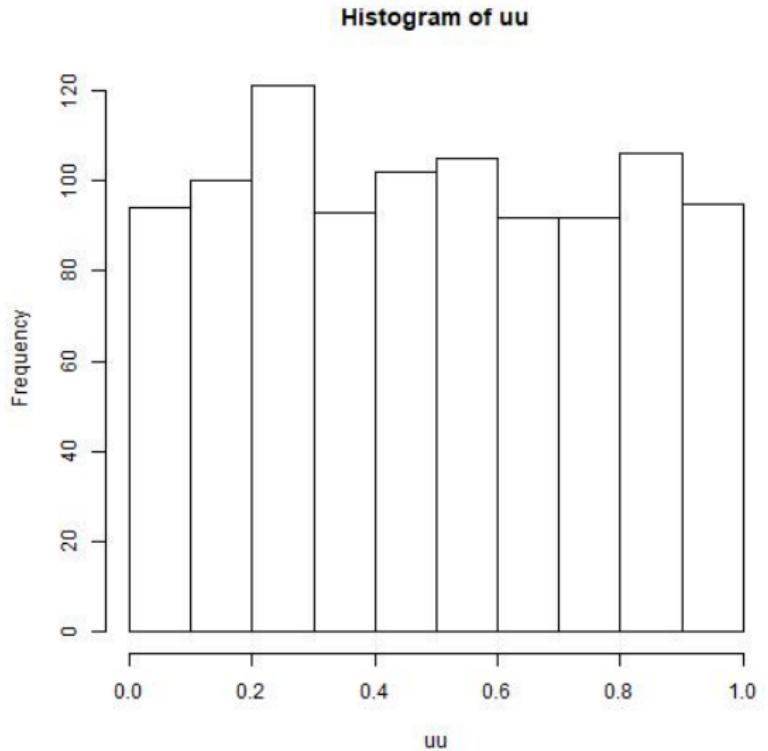
Podríamos listar al vector completo. Si estuvieran bien generados, ¿cuántas coordenadas del **uu** estarían entre 0 y 0.1? ¿Y cuántas entre 0.1 y 0.2?

Cuadro 1: Tabla de frecuencias observadas en el vector uu

intervalo	cantidad
[0, 0,1)	94
[0,1, 0,2)	100
[0,2, 0,3)	121
[0,3, 0,4)	93
[0,4, 0,5)	102
[0,5, 0,6)	105
[0,6, 0,7)	92
[0,7, 0,8)	92
[0,8, 0,9)	106
[0,9, 1)	95

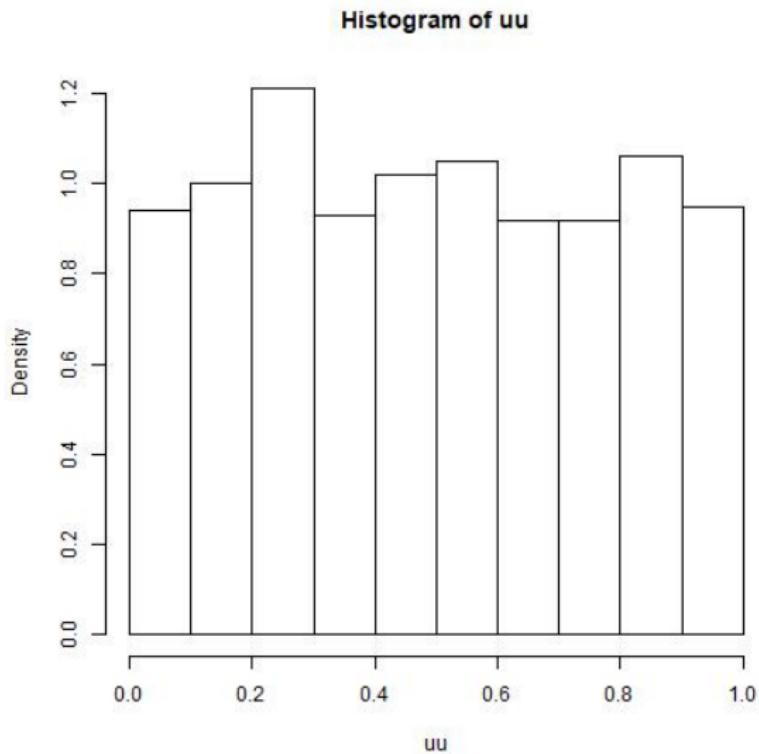
¿Es razonable? ¿Habrá una manera de verlo gráficamente?

Histograma del vector uu



Histograma del vector uu (escala densidad)

(observar la escala del eje y)



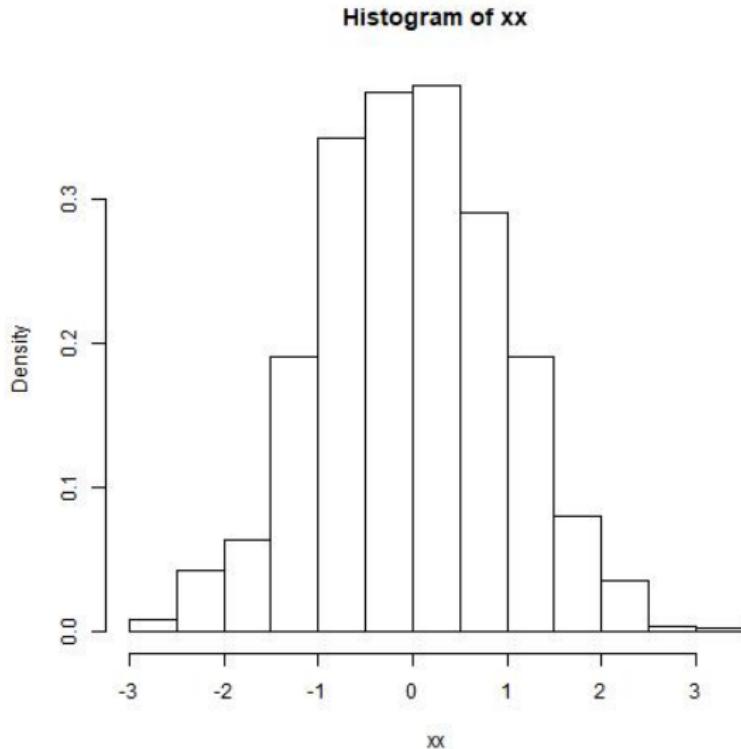
Generamos normales en R usando la proposición

A partir del vector `uu` de longitud 1000, obtenemos otro vector `xx` de longitud 1000, en la coordenada i -ésima de `xx` que llamamos X_i ; tenemos $X_i = \Phi^{-1}(U_i)$. Recordemos que `qnorm` es la instrucción para obtener Φ^{-1} , la inversa de la función de distribución acumulada de una normal estándar. **R** opera fácilmente coordenada a coordenada de un vector. Simplemente tipeamos:

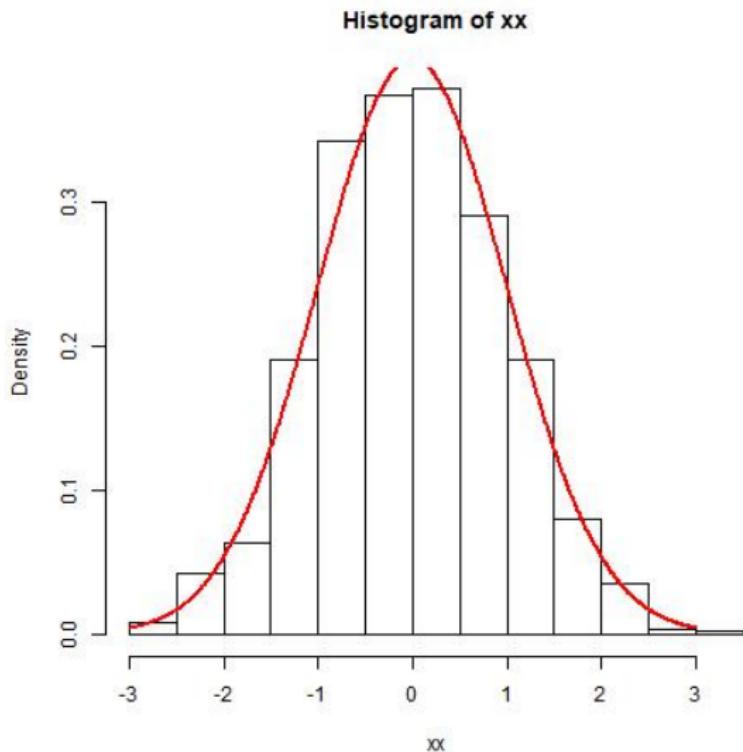
```
xx<-qnorm(uu)
```

¿Cómo será el histograma del vector `xx`? ¿Qué nos dice la Proposición 4.3?

Histograma del vector xx (escala densidad)

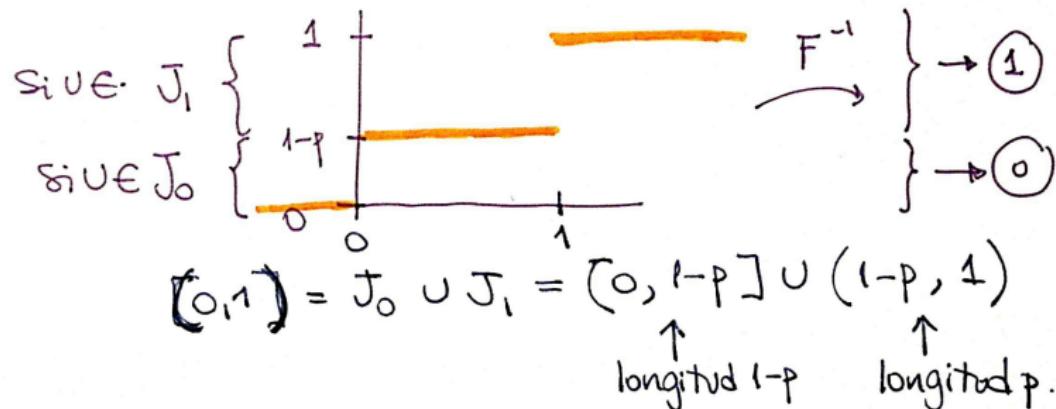


Histograma del vector uu con densidad normal superpuesta



Generar una variable discreta a partir de la $F^{-1}(U)$

¿Cómo haríamos para generar una $X \sim Be(p)$ con este método?

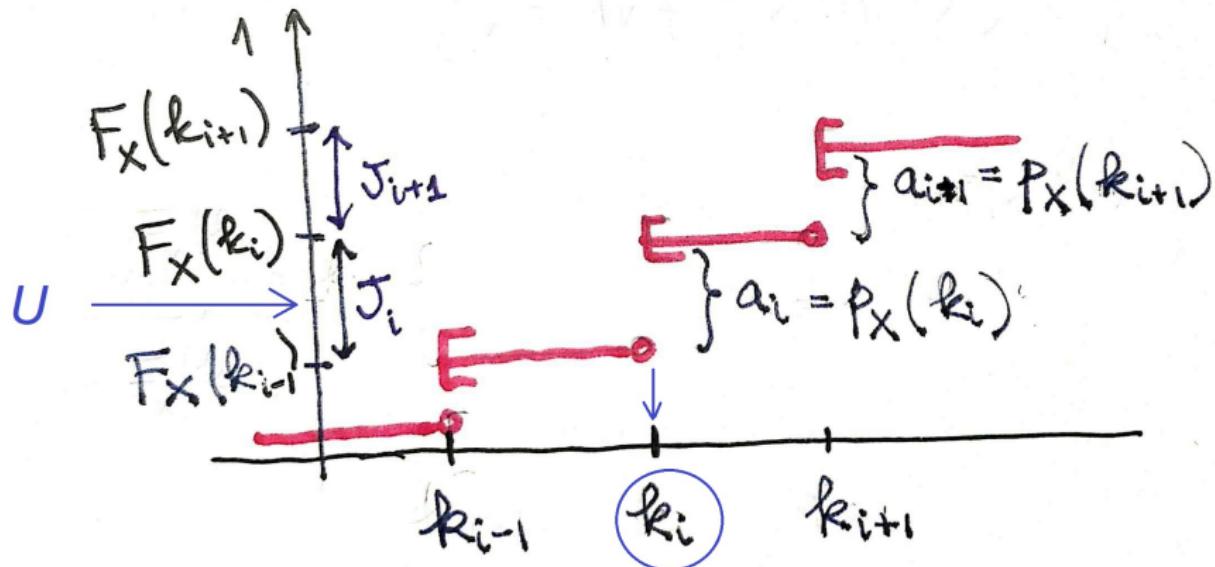


$$P(X=0) = P(F^{-1}(U)=0) = P(U \in \underbrace{(0, 1-p]}_{J_0}) = 1-p.$$

$$P(X=1) = P(U \in J_1) = p.$$

Generar una variable discreta a partir de la $F^{-1}(U)$

¿Cómo haríamos para generar una X discreta con $R_X = \{k_i, i \in \mathbb{N}\}$ y probabilidades puntuales $a_i = p_X(k_i)$ con este método?



Generar una variable discreta a partir de la $F^{-1}(U)$

¿Cómo haríamos para generar una X discreta con $R_X = \{k_i, i \in \mathbb{N}\}$ y probabilidades puntuales $a_i = p_X(k_i)$ con este método? El método consiste en construir una partición del $(0, 1) = \bigcup_{i \in \mathbb{N}} J_i$, con

$$\text{longitud}(J_i) = a_i = p_X(k_i) \quad (10)$$

Si los $\{k_i\}$ estuvieran ordenados en forma creciente (esto no siempre es posible), entonces el método de la inversa generalizada propone elegir $J_i = (F(k_{i-1}), F(k_i)]$. En realidad, basta con que la partición del intervalo $(0, 1)$ cumpla (10). Es decir, si a partir de una v.a. $U \sim \mathcal{U}(0, 1)$ definimos la v.a. X mediante la asignación

$$\{X = k_i\} \Leftrightarrow \{U \in J_i\}$$

con $\{J_i\}_{i \in \mathbb{N}}$ partición que cumple (10) entonces

$$P(X = k_i) = P(U \in J_i) = a_i = p_X(k_i).$$
 😊

La inversa generalizada sólo da una forma de construir esa partición, en el caso de las v.a. discretas.

Variables aleatorias: discretas, continuas y ...

Hasta ahora en la materia solamente nos ocupamos de variables aleatorias discretas y continuas. Hay otra clase de variables aleatorias, denominadas “singulares”. Para una discusión del tema puede consultarse, por ejemplo, Billingsley, P. (1995). *Probability and measure* (3rd edn). Wiley, New York. Un ejemplo usualmente citado de este fenómeno se basa en el conjunto ternario de Cantor. No lo veremos. Puede probarse que una función de distribución variable aleatoria cualquiera se puede escribir como una combinación convexa (una “mezcla” en la terminología probabilística) de variables discreta, continua y singular. En los ejercicios de la materia aparecen ejemplos de combinaciones convexas de continuas y discretas, ya que esta situación es frecuente en las aplicaciones.

Combinación convexa (mezcla) de continua y discreta

Por ejemplo, pensar en la variable $X = \text{"cantidad lluvia caída en una ciudad, en un día elegido al azar"}$. Se puede modelar como un experimento aleatorio que consiste en lanzar una moneda con probabilidad p de obtener cara. Si sale cara (condicional a que sale cara) (si llueve), la cantidad de lluvia caída se modela con una distribución Gama. Si sale ceca (no llueve), la cantidad de lluvia caída es igual a cero. Sea $Y = I_{\text{cara}}$ la indicadora de que la moneda salga cara, $Y \sim Be(p)$, sea $W \sim \Gamma(\alpha, \lambda)$, y Z la variable aleatoria constantemente igual a cero, con función de distribución acumulada $F_Z(x) = I_{[0, +\infty)}(x)$. Entonces,

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(\{X \leq x\} \cap \{Y = 1\}) + P(\{X \leq x\} \cap \{Y = 0\}) \\ &= P(Y = 1)P(X \leq x | Y = 1) + P(Y = 0)P(X \leq x | Y = 0) \\ &= p \cdot P(W \leq x) + (1 - p) \cdot P(Z \leq x) = pF_W(x) + (1 - p)F_Z(x) \end{aligned}$$

Luego, $F_X = p \cdot F_W + (1 - p) \cdot F_Z$, es decir, la función de distribución de X resulta una combinación convexa de una f. de distribución continua (F_W) y una función de distribución discreta (F_Z).

5. Vectores Aleatorios

Probabilidades y Estadística (M)

María Eugenia Szretter Noste

Departamento de Matemática e
Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Primer cuatrimestre 2020



5. Vectores Aleatorios. Repaso de Borelianos

Definición 5.1 (La σ -álgebra de Borel)

Sea $\Omega = \mathbb{R}^k$ y

$$\mathcal{G} = \{[a_1, b_1] \times \cdots \times [a_k, b_k] : a_i \leq b_i, a_i, b_i \in \mathbb{Q}\}$$

la colección que contiene todas las cajas rectangulares en \mathbb{R}^k con vértices racionales y caras paralelas a los ejes. La σ -álgebra $\mathcal{B}^k := \sigma(\mathcal{G}) = \mathcal{B}(\mathbb{R}^k)$ se denomina la σ -álgebra de Borel en \mathbb{R}^k , y a los conjuntos $A \in \mathcal{B}^k = \mathcal{B}(\mathbb{R}^k)$ se los llama conjuntos de Borel o boreelianos. En el caso de $k = 1$, suele escribirse $\mathcal{B}(\mathbb{R}) = \mathcal{B}$ en vez de \mathcal{B}^1 .

tenemos

- a) Todo subconjunto abierto $A \subset \mathbb{R}^k$ es boreiano.
- b) Todo subconjunto cerrado $A \subset \mathbb{R}^k$ es boreiano.
- c) $\mathcal{B}(\mathbb{R}) = \sigma(\{(-\infty, c] : c \in \mathbb{R}\}) = \sigma(\{(a, b] : a \leq b, a, b \in \mathbb{R}\})$

Lema 5.1

Definamos las siguientes familias de conjuntos de \mathbb{R}^k :

- ① $\mathcal{E}_0 = \{[a_1, b_1] \times \cdots \times [a_k, b_k] : a_i \leq b_i, a_i, b_i \in \mathbb{R}, 1 \leq i \leq k\}$.
- ② $\mathcal{E}_1 = \{(a_1, b_1] \times \cdots \times (a_k, b_k] : a_i \leq b_i, a_i, b_i \in \mathbb{R}, 1 \leq i \leq k\}$.
- ③ $\mathcal{E}_2 = \{\text{abiertos en } \mathbb{R}^k\}$.
- ④ $\mathcal{E}_3 = \{\text{bolas en } \mathbb{R}^k\}$
- ⑤ $\mathcal{E}_4 = \{B_1 \times \cdots \times B_k : B_i \in \mathcal{B}(\mathbb{R}), 1 \leq i \leq k\}$.
- ⑥ $\mathcal{E}_5 = \left\{ \bigcup_{i=1}^n \{(a_1^i, b_1^i] \times \cdots \times (a_k^i, b_k^i]\} : a_j^i \leq b_j^i, a_j^i, b_j^i \in \mathbb{R} \right\}$.
- ⑦ $\mathcal{E}_6 = \{(-\infty, a_1] \times \cdots \times (-\infty, a_k] : a_i \in \mathbb{R}\}$

Luego $\mathcal{B}(\mathbb{R}^k) = \sigma(\mathcal{E}_0) = \sigma(\mathcal{E}_1) = \sigma(\mathcal{E}_2) = \sigma(\mathcal{E}_3) = \sigma(\mathcal{E}_4) = \sigma(\mathcal{E}_5) = \sigma(\mathcal{E}_6)$

Probamos $\mathcal{B}(\mathbb{R}^k) = \sigma(\mathcal{E}_0) = \sigma(\mathcal{E}_2) = \sigma(\mathcal{E}_4)$.

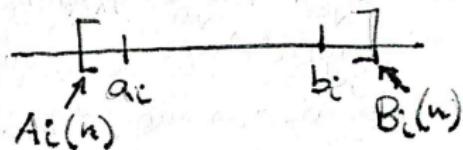
Sea $\mathcal{E}_0 = \{[a_1, b_1] \times \cdots \times [a_k, b_k] : a_i \leq b_i, a_i, b_i \in \mathbb{R}\}$

Luego $\sigma(\mathcal{E}_0) = B(\mathbb{R}^k)$.

dem: Claramente $g \in \mathcal{E}_0 \Rightarrow \sigma(g) = B(\mathbb{R}^k) \subset \sigma(\mathcal{E}_0)$

Sea $[a_1, b_1] \times \cdots \times [a_k, b_k]$ con $a_i, b_i \in \mathbb{R}$. Entonces existen sucesiones

$A_i(n), B_i(n) \in \mathbb{Q}$ tales que



$A_i(n) \nearrow a_i, B_i(n) \searrow b_i$ cuando $n \rightarrow \infty$

Luego $[a_1, b_1] \times \cdots \times [a_k, b_k] = \bigcap_{n=1}^{\infty} ([A_1(n), B_1(n)] \times \cdots \times [A_k(n), B_k(n)]) \in B(\mathbb{R}^k)$

Luego $\mathcal{E}_0 \subseteq \sigma(g) = B(\mathbb{R}^k)$

$\Rightarrow \sigma(\mathcal{E}_0) \subseteq B(\mathbb{R}^k)$ y listo

Sea $\mathcal{E}_2 = \{\text{abiertos de } \mathbb{R}^k\}$

Luego $\sigma(\mathcal{E}_2) = \mathcal{B}(\mathbb{R}^k)$.

dem. Vimos que todo abierto $G \subset \mathbb{R}^k$ es Boreliano,
ya que

$$w \in G \Rightarrow \exists Q_w \in \mathcal{G} / w \in Q_w \subset G$$

$\therefore G = \bigcup_{w \in G} Q_w$ que es una unión

a lo sumo numerable porque hay
numerables elementos en la familia \mathcal{G} .

Luego

$$\mathcal{E}_2 \subset \sigma(\mathcal{G}) = \mathcal{B}(\mathbb{R}^k)$$

$$\Rightarrow \sigma(\mathcal{E}_2) \subseteq \sigma(\mathcal{G}) \quad (\text{A}).$$

Sea F una caja (intervalo) en $\mathcal{G} \Rightarrow F$ es cerrado $\Rightarrow F^c$ abierto

y $F^c = \bigcup_{j \in \mathbb{N}} G_j$ con G_j abiertos $\Rightarrow F = \bigcap_{j \in \mathbb{N}} G_j^c \in \sigma(\mathcal{E}_2)$

$$\Rightarrow \mathcal{G} \subset \sigma(\mathcal{E}_2) \Rightarrow \sigma(\mathcal{G}) \subset \sigma(\mathcal{E}_2) \quad (\text{B})$$

de (A) y (B), listo.

Sea $\mathcal{E}_4 = \{A_1 \times \dots \times A_k : A_i \in \mathcal{B}(\mathbb{R})\}$

Luego $\sigma(\mathcal{E}_4) = \mathcal{B}(\mathbb{R}^k)$.

dem:

\supseteq Sea $(a_1, b_1] \times \dots \times (a_k, b_k] \in \mathcal{E}_4$

Luego $\mathcal{E}_1 \subset \mathcal{E}_4$ y $\sigma(\{(a_1, b_1] \times \dots \times (a_k, b_k] : a_i \leq b_i\})$
 $= \mathcal{B}(\mathbb{R}^k) \subseteq \sigma(\mathcal{E}_4)$

\subseteq

Fijamos $(a_2, b_2] \times \dots \times (a_k, b_k]$

Definimos

$\mathcal{A} = \{A_1 \in \mathcal{B}(\mathbb{R}) : A_1 \times (a_2, b_2] \times \dots \times (a_k, b_k] \in \mathcal{B}(\mathbb{R}^k)\}$

Entonces \mathcal{A} es una colección de subconjuntos Boreelianos de \mathbb{R} . Veamos que:

1) \mathcal{A} es σ -álgebra.

2) Contiene a $\{(a_1, b_1] : a_1 \leq b_1\}$.

Comenzamos por 1)

$$\Rightarrow \emptyset \in \mathcal{A} \text{ pues } \left\{ \emptyset \times [a_2, b_2] \times \dots \times [a_k, b_k] = \emptyset \in \mathcal{B}(\mathbb{R}^k) \right. \\ \left. \emptyset \in \mathcal{B}(\mathbb{R}). \right.$$

$$\Leftrightarrow \text{Sean } (B_j)_{j \in \mathbb{N}} : B_j \in \mathcal{A} \quad \text{y q } \bigcup_{j=1}^{\infty} B_j \in \mathcal{A}$$

$$B_j \in \mathcal{A} \Leftrightarrow \left\{ \begin{array}{l} B_j \in \mathcal{B}(\mathbb{R}) \text{ y} \\ B_j \times [a_2, b_2] \times \dots \times [a_k, b_k] \in \mathcal{B}(\mathbb{R}^k) \end{array} \right.$$

como $\mathcal{B}(\mathbb{R}^k)$ es σ -álgebra, resulta que

$$\bigcup_{j=1}^{\infty} \left(B_j \times [a_2, b_2] \times \dots \times [a_k, b_k] \right) \in \mathcal{B}(\mathbb{R}^k)$$

$$= \left(\bigcup_{j=1}^{\infty} B_j \right) \times [a_2, b_2] \times \dots \times [a_k, b_k]$$

y además $\bigcup_{j=1}^{\infty} B_j \in \mathcal{B}(\mathbb{R})$, luego $\bigcup_{j=1}^{\infty} B_j \in \mathcal{A}$

...> Sea $B \in \mathcal{A}$. Queda $B^c \in \mathcal{A}$

$$B \in \mathcal{A} \Leftrightarrow \begin{cases} B \in \mathcal{B}(\mathbb{R}) \text{ y} \\ B \times [a_2, b_2] \times \cdots \times [a_k, b_k] \in \mathcal{B}(\mathbb{R}^k) \end{cases}$$

Luego $\{B^c \in \mathcal{B}(\mathbb{R})\}$

Para ver que $B^c \in \mathcal{A}$ debemos ver que

$$B^c \times [a_2, b_2] \times \cdots \times [a_k, b_k] \in \mathcal{B}(\mathbb{R}^k)$$

Sabemos que:

$$\underbrace{\mathbb{R} \times [a_2, b_2] \times \cdots \times [a_k, b_k]}_{\prod_{j=2}^k [a_j, b_j]} \in \mathcal{B}(\mathbb{R}^k)$$

Luego

$$B^c \times \prod_{j=2}^k [a_j, b_j] = \underbrace{\left(\mathbb{R} \times \prod_{j=2}^k [a_j, b_j] \right)}_{\in \mathcal{B}(\mathbb{R}^k)} \cap \underbrace{\left(B \times \prod_{j=2}^k [a_j, b_j] \right)}_{\in \mathcal{B}(\mathbb{R}^k)}^c$$

Luego $B^c \in \mathcal{A}$ y $\therefore \mathcal{A}$ resulta σ -álgebra.

Veamos que 2) \mathcal{A} contiene a los intervalos $(a_1, b_1]$
Esto se deduce de que $B(\mathbb{R}^k) = \sigma(E_1)$.

De 1) y 2) resulta que $B(\mathbb{R}) \subseteq \mathcal{A}$ y como $\mathcal{A} \subseteq B(\mathbb{R})$
resulta que $\mathcal{A} = B(\mathbb{R})$.

Luego, acabamos de probar que la colección

$$\mathcal{F}_1 = \left\{ A_1 \times (c_2, d_2] \times \cdots \times (c_k, d_k] : A_1 \in B(\mathbb{R}) \right\} \\ c_i \leq d_i \in \mathbb{R}$$

está contenida en $B(\mathbb{R}^k)$. $\Rightarrow \sigma(\mathcal{F}_1) \subseteq B(\mathbb{R}^k)$.

Por inducción se obtiene el resultado.

Definición 5.2 (Vector aleatorio)

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. $\tilde{\mathbf{X}} : \Omega \rightarrow \mathbb{R}^k$, $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ se dice **vector aleatorio k dimensional** si $\tilde{\mathbf{X}}^{-1}((-\infty, a_1] \times \dots \times (-\infty, a_k]) \in \mathcal{F}$ para todo $a_1, \dots, a_k \in \mathbb{R}$.

Lema 5.2

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad y $\tilde{\mathbf{X}} : \Omega \rightarrow \mathbb{R}^k$ es un vector aleatorio, $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ si y sólo si $\tilde{\mathbf{X}}^{-1}(B) \in \mathcal{F}$ para todo $B \in \mathcal{B}(\mathbb{R}^k)$.

Demostración.

- \Leftarrow) Obvio pues las cajas de la forma $(-\infty, a_1] \times \cdots \times (-\infty, a_k]$ son Boreelianos en \mathbb{R}^k
- \Rightarrow) La familia de conjuntos $\mathcal{A} = \left\{ A \subset \mathbb{R}^k : \tilde{\mathbf{X}}^{-1}(A) \in \mathcal{F} \right\}$ es una σ -álgebra que por hipótesis contiene a los conjuntos $(-\infty, a_1] \times \cdots \times (-\infty, a_k]$, luego contiene a $\mathcal{B}(\mathbb{R}^k)$.



Relación entre vectores y variables aleatorias

Proposición 5.1

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. $\tilde{\mathbf{X}} : (\Omega, \mathcal{F}, P) \rightarrow \mathbb{R}^k$ es un vector aleatorio, $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ si y sólo si cada una de sus coordenadas $X_i : (\Omega, \mathcal{F}, P) \rightarrow \mathbb{R}$ es una variable aleatoria, para todo $i = 1, \dots, k$.

Demostración.

$$\Rightarrow) X_i^{-1}(-\infty, a] = \tilde{\mathbf{X}}^{-1} \left(\mathbb{R} \times \cdots \times \underbrace{(-\infty, a]}_{i\text{-ésima coord}} \times \cdots \times \mathbb{R} \right) \in \mathcal{F}.$$
$$\Leftarrow) \tilde{\mathbf{X}}^{-1} ((-\infty, a_1] \times \cdots \times (-\infty, a_k]) = \bigcap_{i=1}^k X_i^{-1}(-\infty, a_i] \in \mathcal{F}$$



Función de distribución conjunta

Definición 5.3 (Función de distribución conjunta)

La función $F_{\tilde{\mathbf{X}}} : \mathbb{R}^k \rightarrow \mathbb{R}$ es la función de distribución (acumulada) conjunta del vector aleatorio $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ está dada por

$$F_{\tilde{\mathbf{X}}}(x_1, \dots, x_k) = P(X_1 \leq x_1, \dots, X_k \leq x_k)$$

Observemos que la notación es

$$\{X_1 \leq x_1, \dots, X_k \leq x_k\} = \bigcap_{i=1}^k \{X_i \leq x_i\}.$$

Proposición 5.2

$F_{\tilde{\mathbf{X}}}$ identifica la distribución del vector aleatorio $\tilde{\mathbf{X}}$.

Propiedades de la función de distribución conjunta

Proposición 5.3

Sea $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ un vector aleatorio y $F_{\tilde{\mathbf{X}}} = F$ su función de distribución conjunta. Entonces se verifican las siguientes propiedades:

1. $0 \leq F(x_1, \dots, x_k) \leq 1$ para todo $(x_1, \dots, x_k) \in \mathbb{R}^k$.
2. F es una función creciente: si $x_1 \leq y_1, \dots, x_k \leq y_k$ entonces $F(x_1, \dots, x_k) \leq F(y_1, \dots, y_k)$.
3. Si una de las coordenadas converge a $-\infty$, entonces F converge a 0:

$$\lim_{x_i \rightarrow -\infty} F(x_1, \dots, x_k) = 0$$

4. Si todas las coordenadas convergen a $+\infty$, entonces F converge a 1,

$$\lim_{\substack{x_1 \rightarrow +\infty \\ \vdots \\ x_k \rightarrow +\infty}} F(x_1, \dots, x_k) = 1 .$$

Proposición 5.3

5. F es continua a derecha: $x_i^n \searrow x_i$ para todo i , entonces

$$\lim_{n \rightarrow +\infty} F(x_1^n, \dots, x_k^n) = F(x_1, \dots, x_k).$$

6. $P(\tilde{\mathbf{X}} \in (a_1, b_1] \times \dots \times (a_k, b_k]) \geq 0$. Esto se traduce en una condición algebraica. Para $k = 2$, es

$$P(\tilde{\mathbf{X}} \in (a_1, b_1] \times (a_2, b_2]) = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2)$$

$$= \Delta_1(a_1, b_1) \Delta_2(a_2, b_2) F(x_1, x_2) \geq 0,$$

definiendo

$$\Delta_j(a_j, b_j) F_{\tilde{\mathbf{X}}}(\tilde{\mathbf{x}}) =$$

$$F_{\tilde{\mathbf{X}}}(x_1, x_2, \dots, b_j, \dots, x_k) -$$

$$F_{\tilde{\mathbf{X}}}(x_1, x_2, \dots, a_j, \dots, x_k)$$

El caso de k arbitrario,

$$P(\tilde{\mathbf{X}} \in (a_1, b_1] \times \dots \times (a_k, b_k]) = \\ \Delta_1(a_1, b_1) \Delta_2(a_2, b_2) \dots \Delta_k(a_k, b_k) F_{\tilde{\mathbf{X}}}(\tilde{\mathbf{x}}) \geq 0$$

Dem:

- 1) Obvia
- 2) ejercicio

$$3) \{x_1 \leq x_1, \dots, x_i \leq x_i, \dots, x_k \leq x_k\} \subseteq \{x_i \leq x_i\}$$
$$= \bigcap_{j=1}^k \{x_j \leq x_j\}$$

Luego $P(x_1 \leq x_1, \dots, x_k \leq x_k) \leq F_{X_i}(x_i)$

$$F(x_1, \dots, x_k) \leq F_{X_i}(x_i)$$

$$\therefore \lim_{x_i \rightarrow -\infty} F(x_1, \dots, x_k) \leq \lim_{x_i \rightarrow -\infty} F_{X_i}(x_i)$$

$$4) \{X_1 \leq x_1, \dots, X_k \leq x_k\} \nearrow \mathbb{X}^{-1}(\mathbb{R}^k) = \Omega$$

Como el límite existe

(función acotada).

(En cada coordenada es una f. creciente).

(se puede calcular el límite x cualquier sucesión, coinciden)

$$F(x_1, \dots, x_k) = P(X_1 \leq x_1, \dots, X_k \leq x_k)$$

$$\lim_{\substack{x_1, \dots, x_k \rightarrow +\infty}} F(x_1, \dots, x_k) = \lim_{n \rightarrow \infty} F(u, u, \dots, u) = P(\Omega) = 1.$$

$$5) \{X_1 \leq x_1, \dots, X_k \leq x_k\} = \bigcap_{n \geq 1} \{X_1 \leq x_1^n, X_2 \leq x_2^n, \dots, X_k \leq x_k^n\}$$

y el resultado sale de la continuidad
de la probabilidad, pues los A_n son
una sucesión decreciente de eventos

Teorema 5.3

Dada una función de distribución F definida en \mathbb{R}^k satisfaciendo las propiedades 1-6 de la Proposición 5.3, entonces existe un espacio de probabilidad (Ω, \mathcal{F}, P) y un vector aleatorio k -dimensional $\tilde{\mathbf{X}}$ definido en él tal que $F_{\tilde{\mathbf{X}}} = F$.

Distribuciones marginales a partir de la conjunta

A partir de la función de distribución conjunta podemos calcular las funciones de distribución asociadas a cada una de las variables aleatorias que componen al vector aleatorio. Sea $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ un vector aleatorio y $F_{\tilde{\mathbf{X}}}$ su función de distribución conjunta.

$$\begin{aligned} F_{X_i}(t) &= P(X_i \leq t) \\ &= P(X_1 \in \mathbb{R}, X_2 \in \mathbb{R}, \dots, X_i \leq t, X_{i+1} \in \mathbb{R}, \dots, X_k \in \mathbb{R}) \\ &= \lim_{\substack{x_1 \rightarrow +\infty \\ x_2 \rightarrow +\infty \\ \vdots \\ x_{i-1} \rightarrow +\infty \\ x_{i+1} \rightarrow +\infty \\ \vdots \\ x_k \rightarrow +\infty}} F_{\tilde{\mathbf{X}}}(x_1, x_2, \dots, x_{i-1}, t, x_{i+1}, \dots, x_k). \end{aligned}$$

Cuando hablamos de la función de distribución de una variable aleatoria en el contexto de vectores aleatorios, se suele enfatizar el hecho de que es univariada hablando de la **función de distribución marginal** de X_i .

Distribuciones de subvectores aleatorios

A partir de la función de distribución conjunta también podemos calcular la función de distribución asociada a un subvector del vector aleatorio. Sea $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ un vector aleatorio y $F_{\tilde{\mathbf{X}}}$ su función de distribución conjunta. Podemos dividir al vector $\tilde{\mathbf{X}} = (\tilde{\mathbf{U}}, \tilde{\mathbf{V}})$ de modo que $\tilde{\mathbf{U}} \in \mathbb{R}^{k_1}$ y $\tilde{\mathbf{V}} \in \mathbb{R}^{k_2}$ con $k_1 + k_2 = k$. Por la Proposición 5.1 sabemos que tanto $\tilde{\mathbf{U}}$ como $\tilde{\mathbf{V}}$ son vectores aleatorios. Podemos obtener sus funciones de distribución conjunta a partir de $F_{\tilde{\mathbf{X}}}$.

$$\begin{aligned} F_{\tilde{\mathbf{U}}}(\mathbf{u}) &= P\left(\bigcap_{i=1}^{k_1} \{X_i \leq u_i\}\right) = P\left(\bigcap_{i=1}^{k_1} \{X_i \leq u_i\} \cap \bigcap_{i=k_1+1}^k \{X_i \in \mathbb{R}\}\right) \\ &= \lim_{\substack{x_{k_1+1} \rightarrow +\infty \\ \vdots \\ x_k \rightarrow +\infty}} F_{\tilde{\mathbf{X}}}(\color{red}{u_1, u_2, \dots, u_{k_1}, x_{k_1+1}, \dots, x_k}). \end{aligned}$$

Ejemplos de vectores aleatorios

Ejemplo 5.1 (Bolitas de colores en una urna)

Se tiene una urna con 20 bolas rojas, 15 bolas azules, 10 bolas verdes y 5 bolas anaranjadas. Se extraen 8 bolas sin reposición. Interesa estudiar la distribución del vector aleatorio $\tilde{\mathbf{X}} = (X_1, X_2, X_3, X_4)$, donde cada variable cuenta la cantidad de bolas del color correspondiente que fueron extraídas entre las 8 seleccionadas. Luego, por ejemplo,

$$P(\tilde{\mathbf{X}} = (5, 2, 1, 0)) = \frac{\binom{20}{5} \binom{15}{2} \binom{10}{1} \binom{5}{0}}{\binom{50}{8}} = 0.0303$$

Ejemplo 5.2 (Variables antropométricas)

Interesa estudiar la relación entre distintas variables aleatorias medidas sobre el mismo individuo. Para ello, a 246 hombres adultos elegidos al azar se les midieron varias variables antropométricas (edad, peso, altura, contorno de tobillos, de abdomen, de cuello, etc.) ¿Qué vínculos tienen entre sí estas variables? Si se arman patrones que reflejen el vínculo entre ellas en poblaciones sanas, se puede contribuir al diagnóstico del estado de salud, por ejemplo. Además, algunas medidas muy precisas del porcentaje de grasa corporal son bastante difíciles de obtener (involucran medir la cantidad de líquido desplazado al sumergirse en una bañera, por ejemplo) y su valor en un individuo se podría predecir usando un modelo matemático calibrado con las otras mediciones.

Referencia: Penrose, K., Nelson, A., and Fisher, A. (1985), "Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques" (abstract), *Medicine and Science in Sports and Exercise*, 17(2), 189. y datos analizados por Johnson, R.W. Fitting Percentage of Body Fat to Simple Body Measurements, *Journal of Statistics Education* v.4, n.1 (1996)

Datos en <http://jse.amstat.org/v4n1/datasets.johnson.html>

Ejemplo 5.2 (Variables antropométricas, cont.)

Definimos las variables

X_1 = edad (años)

X_2 = altura (cm.)

X_3 = peso (kg.)

X_4 = contorno del cuello (cm.)

X_5 = contorno del pecho (cm.)

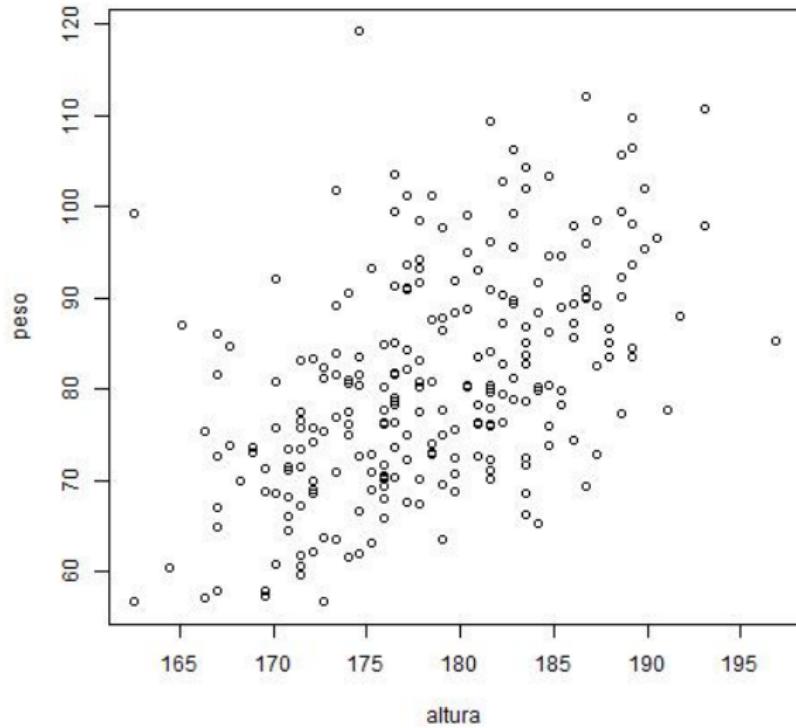
X_6 = contorno del abdomen (cm.)

X_7 = contorno del tobillo (cm.)

Nos interesa el vector aleatorio $\tilde{\mathbf{X}} = (X_1, \dots, X_7)$. Para darnos una idea de su comportamiento univariado (variable por variable) podemos hacer un histograma de cada variable, basado en las 246 observaciones. Para entender el comportamiento conjunto podemos hacer un gráfico en \mathbb{R}^2 de las observaciones, tomadas de a pares de variables. Se suelen llamar *gráficos de dispersión* o *scatter plots*. Veamos algunos.

Variables antropométricas: peso vs. altura

¿Qué graficamos en el scatter plot?

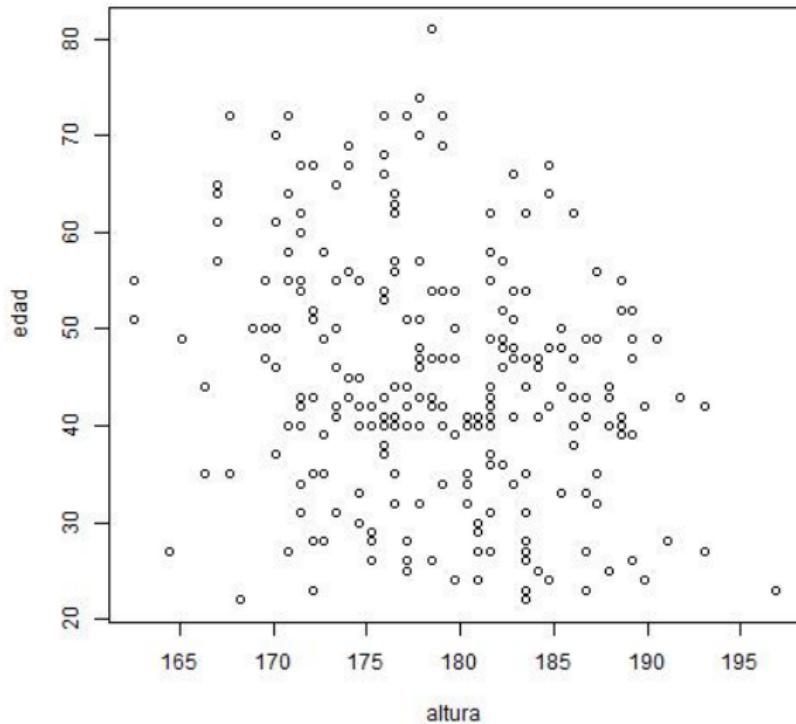


- cada punto es un individuo, graficamos el par $(X_2, X_3)(\omega) = (\text{altura}, \text{peso})(\omega)$ con (ω) un paciente
- por ej (175,120), ¿qué significa?
- perdemos la identificación del paciente

¿Qué vemos en este gráfico?

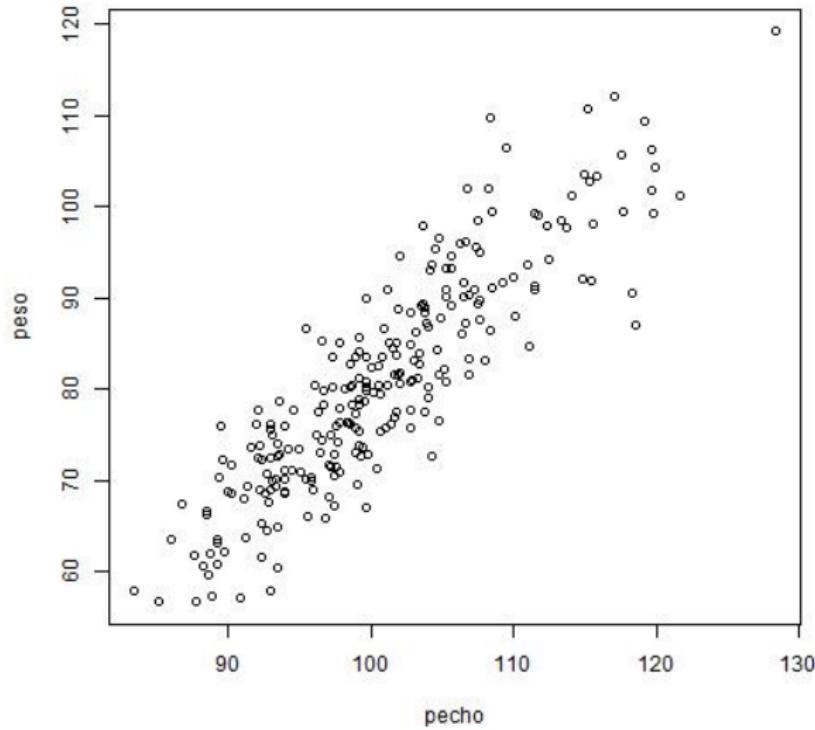
- relación creciente (a mayor altura hay, en general, mayor peso)
- bastante dispersión
- datos acomodados de forma elíptica

Variables antropométricas: edad vs. altura



- no hay relación entre ambas variables
- mucha dispersión
- datos acomodados de forma circular

Variables antropométricas: contorno del pecho vs. peso



- fuerte relación creciente entre ambas variables
- menos dispersión, una recta resume bastante bien el vínculo
- datos acomodados de forma elíptica, pero una elipse “achatada”

Ejemplo 5.3 (Temperatura máxima y mínima, en Argentina)

La página del Servicio Meteorológico Nacional permite el libre acceso a los datos de temperatura máxima y mínima diarias registradas durante el último año en 123 estaciones meteorológicas, varias de las cuales están ubicadas en los aeropuertos de las ciudades. Hay datos disponibles de precipitaciones, velocidad del viento, radiación solar, presión, y otras más.

<https://www.smn.gob.ar/descarga-de-datos>

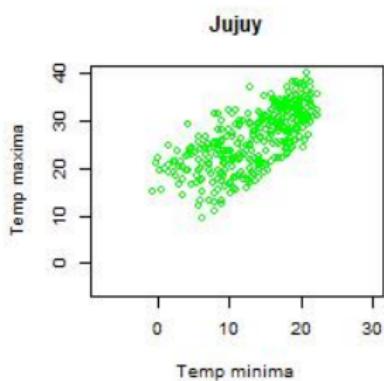
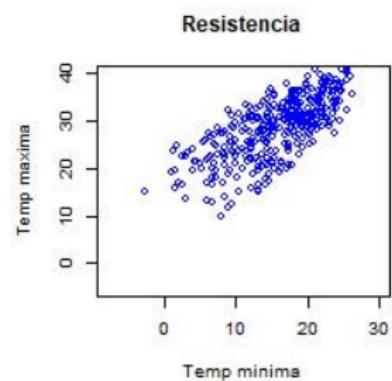
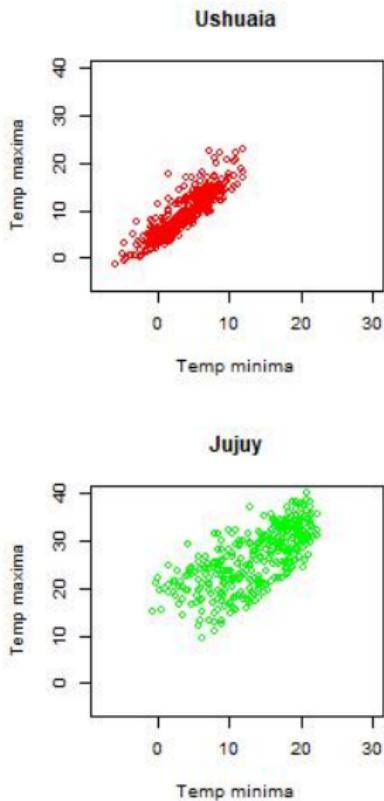
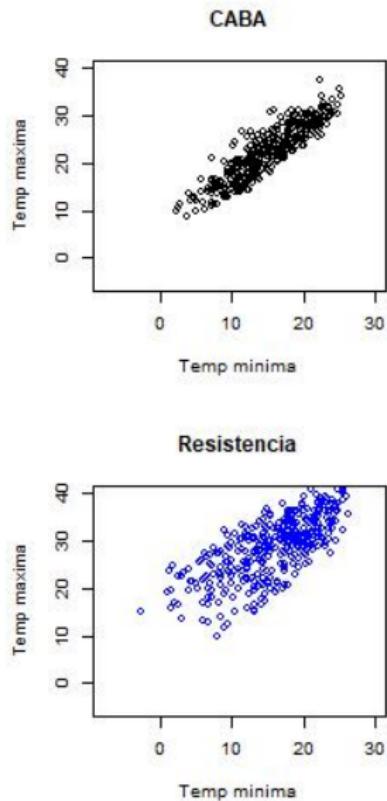
Como ejemplo de vectores aleatorios podemos considerar el vector

- (X_1, X_2) que consiste en registrar la temperatura mínima y máxima de la Ciudad Autónoma de Buenos Aires, durante un día determinado del último año

Y también

- (W_1, W_2) , temperatura mínima y máxima de Ushuaia.
- (Y_1, Y_2) , temperatura mínima y máxima de Resistencia.
- (Z_1, Z_2) , temperatura mínima y máxima de Jujuy.

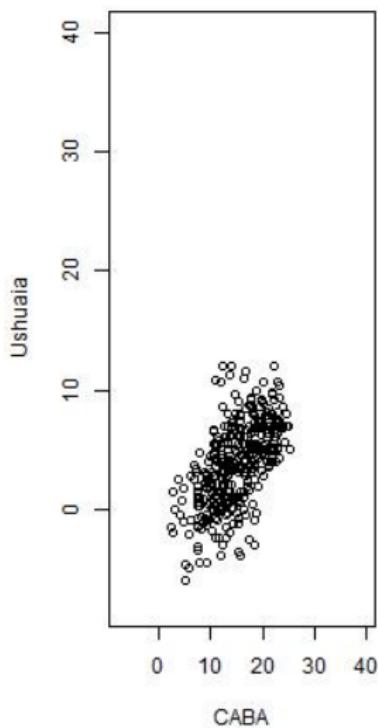
Temperatura máxima y mínima diaria, por localidad



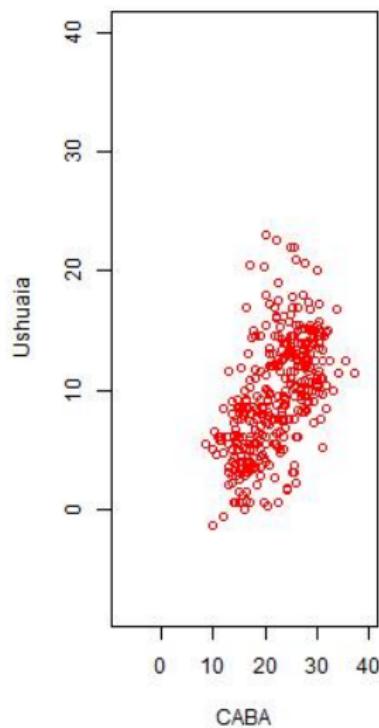
- hay relación creciente entre ambas variables
- en algunos, la asociación es más fuerte
- con distinta dispersión o variabilidad en cada distrito
- datos no necesariamente elípticos

Temperaturas mínimas (máximas) CABA y Ushuaia

Temperatura minima



Temperatura maxima



- primer gráfico es (X_1, W_1) , el segundo (X_2, W_2) para cada día
- hay relación creciente entre ambas variables
- pueden estudiarse patrones geográficos
- pueden estudiarse patrones temporales
- estos gráficos son descriptivos

Ejemplo 5.4 (cine)

Alicia y José acordaron encontrarse a las 8 de la noche para ir al cine. Como no son puntuales, se puede suponer que los tiempos X e Y en que cada uno de ellos llega son variables aleatorias con distribución uniforme entre las 8 y las 9. Además se supondrá que estos tiempos son independientes. Ambos están dispuestos a esperar al otro no más de 10 minutos a partir del instante en que llegan, ¿cuál es la probabilidad de que se encuentren? Si bien en este caso entendemos la distribución de X e Y por separado, para poder responder bien a la pregunta debemos ser capaces de calcular:

$$P(|X - Y| < 1/6) = P((X, Y) \in C)$$

donde C es un subconjunto de \mathbb{R}^2 .

Independencia de Variables Aleatorias

Definición 5.4 (variables independientes)

Una familia $\{X_i\}_{i \in \mathbb{N}}$ de variables aleatorias definidas en el mismo espacio de probabilidad (Ω, \mathcal{F}, P) se dice independiente si y sólo si

$$P\left(\bigcap_{i=1}^k X_i^{-1}(B_i)\right) = \prod_{i=1}^k P(X_i^{-1}(B_i)), \quad (1)$$

para toda elección $B_i \in \mathcal{B}$ y para todo $k \in \mathbb{N}$.

Recordemos que también notábamos

$$P\left(\bigcap_{i=1}^k X_i^{-1}(B_i)\right) = P(X_1 \in B_1, \dots, X_k \in B_k)$$

Es decir que los eventos $\{X_1 \in B_1\}, \dots, \{X_k \in B_k\}$ son independientes

En realidad, la definición de independencia de eventos que dimos en el Capítulo 2, involucraba que la igualdad (1) se cumpliera también cuando suprimimos algunos eventos de la forma $\{X_i \in B_i\}$ de cada lado, pero esto sólo significa tomar $B_i = \mathbb{R}$.

Proposición 5.4

X_1, \dots, X_k definidas en (Ω, \mathcal{F}, P) son independientes si y sólo si la función de distribución acumulada conjunta se factoriza, es decir,

$$F_{\tilde{\mathbf{X}}}(x_1, x_2, \dots, x_k) = \prod_{i=1}^k F_{X_i}(x_i),$$

con $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$, para todo $(x_1, x_2, \dots, x_k) \in \mathbb{R}^k$.

Sabemos que

$$F_{xy}(x,y) = F_x(x) F_y(y) \quad \forall x,y \in \mathbb{R}.$$

O equivalentemente que

$$P(x \in (-\infty, x], y \in (-\infty, y]) = P(x \in (-\infty, x]) P(y \in (-\infty, y]) \quad \forall x, y \in \mathbb{R} \quad (1)$$

Queremos ver que

$$P(x \in A, y \in B) = P(x \in A) \cdot P(y \in B) \quad \forall A, B \in \mathcal{B}(\mathbb{R}) \quad (2)$$

La diferencia entre (1) y (2) es el cuantificador.

Dicho de otro modo, sabemos que las 2 probabilidades dadas por

- la conjunta
- el producto de las marginales

coinciden en los conjuntos de \mathbb{R}^2 de la forma

$$J_i = \{(-\infty, x] \times (-\infty, y] : x, y \in \mathbb{R}\}$$

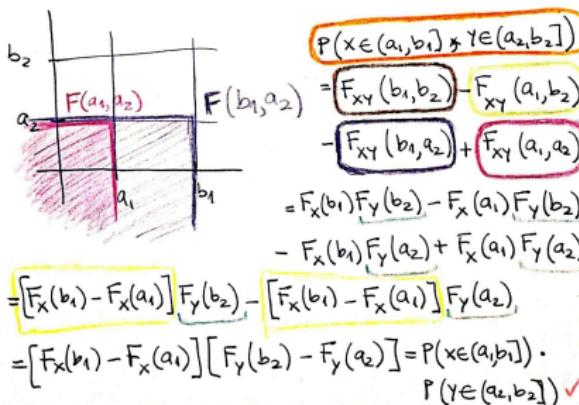
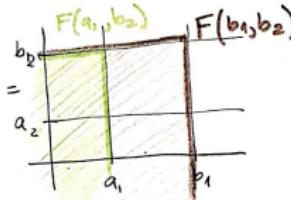
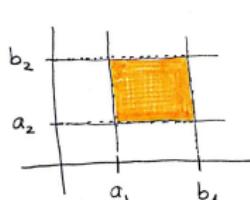
y queremos probar que valen en los conjuntos de la forma:

$$A = \{A \times B : A, B \in \mathcal{B}(\mathbb{R})\}.$$

Esta extensión puede hacerse (muuy trabajosamente) "ad-hoc" o usando teoría más general, que excede el contenido de esta materia.

Veamos cómo extenderla a la familia

$$\mathcal{G}_2 = \{(a_1, b_1] \times (a_2, b_2] : a_1 \leq b_1, a_2 \leq b_2\}$$



La demostración del caso general se parece a la demostración de que la σ -álgebra de Borel en \mathbb{R}^k está generada por producto de Borelianos en \mathbb{R} . Si la quieren leer, el libro es una buena fuente

Walsh, J. B. (2012) Knowing the odds: an Introduction to Probability (Graduate Studies in Mathematics, Vol 139).

Definición 5.5

$\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_k$ vectores aleatorios definidos en (Ω, \mathcal{F}, P) y tomando valores en \mathbb{R}^{n_i} , para $1 \leq i \leq k$, se dicen independientes si y sólo si

$$P\left(\bigcap_{i=1}^k \tilde{\mathbf{X}}_i^{-1}(B_i)\right) = \prod_{i=1}^k P(\tilde{\mathbf{X}}_i^{-1}(B_i)),$$

para toda elección $B_i \in \mathcal{B}^{n_i}$.

Del mismo modo que tenemos la Proposición 5.4 para variables aleatorias, para vectores aleatorios se tiene también la equivalencia con el hecho de que la función de distribución conjunta se factoriza en k factores (de funciones de distribución **conjunta** n_i dimensional cada uno).

Diremos que una función $g : \mathbb{R}^k \rightarrow \mathbb{R}^j$ es **boreiana** (o medible Borel) si para todo $A \in \mathcal{B}(\mathbb{R}^j)$ vale que $g^{-1}(A) \in \mathcal{B}(\mathbb{R}^k)$.

Lema 5.4

Sea $g : \mathbb{R}^k \rightarrow \mathbb{R}^l$ una función continua, entonces es medible Borel. Luego, si $\tilde{\mathbf{X}}$ es un vector aleatorio k -dimensional, tenemos que $\tilde{\mathbf{Y}} = g(\tilde{\mathbf{X}})$ es un vector aleatorio l dimensional.

Demostración.

Ejercicio. □

Lema 5.5 (la independencia se mantiene por transformaciones)

Sean $g_i : \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{l_i}$ boreianas. $\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_k$ vectores aleatorios independientes. Entonces, los vectores $\tilde{\mathbf{Y}}_1 = g_1(\tilde{\mathbf{X}}_1), \dots, \tilde{\mathbf{Y}}_k = g_k(\tilde{\mathbf{X}}_k)$ también son independientes.

Demostración.

Aplicamos la definición de independencia. Dados B_1, B_2, \dots, B_k boreianos arbitrarios de $\mathbb{R}^{l_1}, \dots, \mathbb{R}^{l_k}$ queremos probar que los conjuntos

$$\tilde{\mathbf{Y}}_1^{-1}(B_1), \tilde{\mathbf{Y}}_2^{-1}(B_2), \dots, \tilde{\mathbf{Y}}_k^{-1}(B_k)$$

son eventos independientes. Ahora bien para cada $j = 1, 2, \dots, k$ se tiene

$$\tilde{\mathbf{Y}}_j^{-1}(B_j) = \tilde{\mathbf{X}}_j^{-1}\left(g_j^{-1}(B_j)\right)$$

Como los $g_j^{-1}(B_j), j = 1, 2, \dots, k$ son boreianos porque las g_j lo son, la independencia de los vectores $\tilde{\mathbf{X}}_j$ implica que los eventos $\tilde{\mathbf{Y}}_j^{-1}(B_j)$ son independientes, probando el resultado. □

Las funciones de distribución conjunta identifican a la distribución marginal, pero no vale al revés. Veremos contraejemplos. Es decir, dada la función de distribución conjunta de un vector (X, Y) , F_{XY} podemos hallar las funciones de distribución (marginal) de cada variable F_X y F_Y (de forma única),

$$F_X(x) = \lim_{y \rightarrow +\infty} F_{XY}(x, y)$$

Pero al revés, dados F_X y F_Y hay muchas F_{XY} compatibles con estas distribuciones marginales.

Vectores Discretos

Definición 5.6 (vectores discretos)

$\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ se dice un vector aleatorio discreto si cada una de las coordenadas es una variable aleatoria discreta.

Recordemos que como X_i es discreta para cada i , el rango

$$R_{X_i} = \{k \in \mathbb{R} : p_{X_i}(k) > 0\}$$

es un conjunto a lo sumo numerable.

En tal caso, tendríamos que $P(X_1 \in R_{X_1}, \dots, X_k \in R_{X_k}) = 1$ con lo que $P(\tilde{\mathbf{X}} \in R_{X_1} \times \dots \times R_{X_k}) = 1$. Como producto finito de numerables es numerable, si denotamos por $A = R_{X_1} \times \dots \times R_{X_k}$ tenemos que $P(\tilde{\mathbf{X}} \in A) = 1$ para $A \subseteq \mathbb{R}^k$ numerable. Definimos el

rango del vector $\tilde{\mathbf{X}}$ discreto por

$$R_{\tilde{\mathbf{X}}} = \left\{ \tilde{\mathbf{x}} \in \mathbb{R}^k : P(\tilde{\mathbf{X}} = \tilde{\mathbf{x}}) > 0 \right\}$$

Resulta $R_{\tilde{\mathbf{X}}} \subset A$ y por lo tanto es a lo sumo numerable.

Recordemos que $\{\tilde{\mathbf{X}} = \tilde{\mathbf{x}}\} = \{X_1 = x_1\} \cap \dots \cap \{X_k = x_k\}$

Función de probabilidad puntual conjunta

Dado un vector aleatorio discreto $\tilde{\mathbf{X}}$ definimos su **función de probabilidad puntual conjunta** haciendo

$$p_{\tilde{\mathbf{X}}}(x_1, \dots, x_k) = P(\tilde{\mathbf{X}} = (x_1, \dots, x_k)).$$

Entonces, como en el caso de variables discretas,

$$P(\tilde{\mathbf{X}} \in C) = \sum_{\tilde{\mathbf{x}} \in C \cap R_{\tilde{\mathbf{X}}}} p_{\tilde{\mathbf{X}}}(\tilde{\mathbf{x}})$$

con $\tilde{\mathbf{x}} = (x_1, \dots, x_k)$ y cualquier $C \in \mathcal{B}(\mathbb{R}^k)$.

Resulta

$$\sum_{(x_1, \dots, x_k) \in R_{\tilde{\mathbf{X}}}} p_{\tilde{\mathbf{X}}}(x_1, \dots, x_k) = 1.$$

Función de Probabilidad Puntual Marginal

Podemos calcular la **función de probabilidad puntual marginal** de cada X_i :

$$p_{X_i}(\textcolor{red}{x}_i) = \sum_{\substack{\{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k : \\ (x_1, \dots, x_{i-1}, \textcolor{red}{x}_i, x_{i+1}, \dots, x_k) \in R_{\tilde{\mathbf{X}}}\}}} p_{\tilde{\mathbf{X}}}(x_1, \dots, x_{i-1}, \textcolor{red}{x}_i, x_{i+1}, \dots, x_k).$$

Podemos escribir a la función de distribución conjunta acumulada sumando puntuales.

$$F_{\tilde{\mathbf{X}}}(x_1, x_2, \dots, x_k) = \sum_{\substack{(y_1, \dots, y_k) \in R_{\tilde{\mathbf{X}}}: \\ y_1 \leq x_1, \dots, y_k \leq x_k}} p_{\tilde{\mathbf{X}}}(y_1, \dots, y_k)$$

Proposición 5.5

Sean $X_1, \dots, X_k : \Omega \rightarrow \mathbb{R}$ variables aleatorias discretas. Son equivalentes:

① X_1, \dots, X_k son independientes.

② la probabilidad puntual conjunta se factoriza:

$$p_{X_1, \dots, X_k}(x_1, \dots, x_k) = p_{X_1}(x_1) \cdots p_{X_k}(x_k), \text{ para todo } x_1, \dots, x_k.$$

③ la función de distribución conjunta se factoriza:

$$F_{X_1, \dots, X_k}(x_1, \dots, x_k) = F_{X_1}(x_1) \cdots F_{X_k}(x_k), \text{ para todo } x_1, \dots, x_k \in \mathbb{R}.$$

④ Existen funciones $h_i : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ tales que

$$p_{X_1, \dots, X_k}(x_1, \dots, x_k) = h_1(x_1) \cdots h_k(x_k), \text{ para todo } x_1, \dots, x_k \in \mathbb{R}.$$

⑤ Existen funciones $H_i : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ tales que

$$F_{X_1, \dots, X_k}(x_1, \dots, x_k) = H_1(x_1) \cdots H_k(x_k), \text{ para todo } x_1, \dots, x_k \in \mathbb{R}.$$

dem:

1) \Leftrightarrow 3) es la proposición anterior.

1) \Rightarrow 2) Fácil.

2) \Rightarrow 1)

$$P(X_1 \in B_1, \dots, X_k \in B_k) = \sum_{(x_1, \dots, x_k)} p_X(x_1, \dots, x_k) = \text{(*)}$$

$$(x_1, \dots, x_k) \in R_X \cap B_1 \times \dots \times B_k$$

Queremos poner ahora $R_{X_1} \times \dots \times R_{X_k}$ en vez de R_X . Notemos que por lo comentado antes de definir el rango de X tenemos que:

$$R_X \subset R_{X_1} \times \dots \times R_{X_k}. \text{ Más aún } R_{X_1} \times \dots \times R_{X_k} = R_X \cup C \quad (\text{disjunta})$$

$$\text{y si } y \in C \text{ se cumple que } p_X(y) = 0. \quad (C = R_{X_1} \times \dots \times R_{X_k} - R_X)$$

Luego, podemos escribir la suma de (*) así:

$$\begin{aligned} (*) &= \sum_{(x_1, \dots, x_k)} p_X(x_1, \dots, x_k) = \sum_{(x_1, \dots, x_k)} p_{X_1}(x_1) \dots p_{X_k}(x_k) \\ &\quad (x_1, \dots, x_k) \in R_{X_1} \times \dots \times R_{X_k} \cap B_1 \times \dots \times B_k \quad (x_1, \dots, x_k) \in (R_{X_1} \cap B_1) \times \dots \times (R_{X_k} \cap B_k) \\ &= \left[\sum_{x_1 \in B_1 \cap R_{X_1}} p_{X_1}(x_1) \right] \dots \left[\sum_{x_k \in R_{X_k} \cap B_k} p_{X_k}(x_k) \right] = \prod_{i=1}^k P(X_i \in B_i) \end{aligned}$$

2) \Rightarrow 4) fácil, tomar $h_i = p_{X_i}$

4) \Rightarrow 2) $p_{X_i}(x_i) = \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k} p_{x_1, \dots, x_k}(x_1, \dots, x_{i-1}, \underset{\text{fijo}}{\overset{x_i}{\uparrow}}, x_{i+1}, \dots, x_k)$

(no sumamos sobre la
íesima coordenada:
queda fija en x_i)

$$= \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k} q_1(x_1) \cdots q_{i-1}(x_{i-1}) \underset{\text{q. fija}}{\underset{x_i}{\text{q}_i(x_i)}} q_{i+1}(x_{i+1}) \cdots q_k(x_k)$$

$$= \underset{\text{q. fija}}{\underset{x_i}{\text{q}_i(x_i)}} \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k} q_1(x_1) \cdots q_{i-1}(x_{i-1}) q_{i+1}(x_{i+1}) \cdots q_k(x_k)$$

↑ es una suma fija. (constante que no depende de x_i). La llamamos c_i

se suma sobre todos los x_j ($j \neq i$) tales que $(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_k) \in R_{\tilde{x}}$.

Luego:

$$p_{X_i}(x_i) = c_i q_i(x_i)$$

Luego:

$$p_X(x_1, \dots, x_k) = \prod_{i=1}^k q_i(x_i) = \frac{1}{\prod_{i=1}^k c_i} \frac{p_{X_i}(x_i)}{c_i} = \frac{\prod_{i=1}^k p_{X_i}(x_i)}{\prod_{i=1}^k c_i}$$

Pero como $1 = \sum_{x_1, \dots, x_k} p_X(x_1, \dots, x_k) = \frac{1}{\prod_{i=1}^k c_i} \sum_{j=1}^k p_{X_1}(x_1) \dots p_{X_k}(x_k)$

$$= \frac{1}{\prod_{i=1}^k c_i} \left[\sum_{x_1} p_{X_1}(x_1) \right] \dots \left[\sum_{x_k} p_{X_k}(x_k) \right] = \frac{1}{\prod_{i=1}^k c_i}$$

Resulta
 $\boxed{\prod_{j=1}^k c_j = 1}$
y vale 2)

3) \Rightarrow 5) fácil, basta tomar $H_i = F_{X_i}$.

5) \Rightarrow 3). Para simplificar la escritura trabajemos con $i=1$.

$$F_{X_1}(\underline{x}_1) = \lim_{\substack{x_2 \rightarrow +\infty \\ \vdots \\ x_k \rightarrow +\infty}} F_X(\underline{x}, \dots, x_k) . \underset{\substack{\uparrow \\ \text{fijo}}}{=} \lim_{\substack{x_2 \rightarrow +\infty \\ \vdots \\ x_k \rightarrow +\infty}} H_1(x_1) \prod_{j=2}^k H_j(x_j)$$

$\xrightarrow{\text{x hipótesis:}}$

$$= H_1(\underline{x}_1) \left[\prod_{j=2}^k \lim_{x_j \rightarrow +\infty} H_j(x_j) \right]$$

D_1 (no depende de \underline{x}_1).

$$\text{Luego } F_{X_1}(\underline{x}_1) = D_1 H_1(\underline{x}_1) \quad \text{y} \quad F_{X_j}(\underline{x}_j) = D_j H_j(\underline{x}_j)$$

También:

$$1 = \lim_{\substack{x_1 \rightarrow +\infty \\ \vdots \\ x_k \rightarrow +\infty}} F_{X_1 \dots X_k}(x_1, \dots, x_k) = \lim_{\substack{x_1 \rightarrow +\infty \\ \vdots \\ x_k \rightarrow +\infty}} \prod_{j=1}^k H_j(x_j)$$

$$= \lim_{\substack{x_1 \rightarrow +\infty \\ \vdots \\ x_k \rightarrow +\infty}} \prod_{j=1}^k \frac{F_{X_j}(x_j)}{D_j} = \frac{1}{\prod_{j=1}^k D_j} \prod_{j=1}^k \underbrace{\lim_{x_j \rightarrow +\infty} F_{X_j}(x_j)}_1$$

$$\Rightarrow 1 = \prod_{j=1}^k D_j. \text{ Luego}$$

$$\begin{aligned} F_{X_1 \dots X_k}(x_1, \dots, x_k) &= \prod_{j=1}^k H_j(x_j) = \frac{\prod_{j=1}^k F_{X_j}(x_j)}{\prod_{l=1}^k D_l} \\ &= \prod_{j=1}^k F_{X_j}(x_j) \quad \checkmark \end{aligned}$$

Las equivalencias 4) y 5) de la Proposición dicen que basta factorizar la probabilidad puntual conjunta o la función de distribución conjunta, aún cuando los factores no sean funciones de probabilidad puntual o funciones de distribución (de hecho, lo que les faltará es una constante multiplicativa), basta con que sean mayores o iguales que cero.

Ejercicio 5.1 (Salario)

En una cierta población, se elige un trabajador mayor de 30 años. Sean

X = cantidad de años de educación que recibió

Y = salario que cobra (en miles de pesos)

Se sabe que la función de probabilidad puntual del vector aleatorio (X, Y) está dado por $p_{XY}(x, y)$

		X			
		7	12	18	24
Y	40	0.14	0.23	0.02	0.01
	100	0.06	0.16	0.25	0.03
	150	0	0.01	0.03	0.06

(es decir, $0.23 = p_{XY}(12, 40)$).

- ① Hallar p_X y p_Y . ¿Son las variables X e Y independientes?
- ② Suponga que las variables X e Y fueran independientes, con las funciones de probabilidad puntual que calculó en el ítem 1). Halle la probabilidad conjunta en este caso y compárela con la dada.

Ejercicio 5.1, (Salario, respuesta)

Las probabilidades puntuales marginales resultan ser (tabla original)

		X				$p_Y(y)$
y \ x		7	12	18	24	
Y	40	0.14	0.23	0.02	0.01	0.40
	100	0.06	0.16	0.25	0.03	0.50
	150	0	0.01	0.03	0.06	0.10
$p_X(x)$		0.20	0.40	0.30	0.10	1

Si X e Y fueran independientes, la conjunta $p_{XY}(x, y) = p_X(x)p_Y(y)$

		X				$p_Y(y)$
y \ x		7	12	18	24	
Y	40	0.08	0.16	0.12	0.04	0.40
	100	0.10	0.20	0.15	0.05	0.50
	150	0.02	0.04	0.03	0.01	0.10
$p_X(x)$		0.20	0.40	0.30	0.10	1

Vector discreto famoso: Distribución Multinomial

Repetimos n veces de manera independiente un experimento que tiene k resultados posibles. El resultado i -ésimo tiene una probabilidad p_i de ser obtenido en cada realización del experimento ($i = 1, \dots, k$). Luego, para cada $i = 1, \dots, k$ definimos $X_i =$ número de veces en las que se obtuvo el resultado i , de las n repeticiones. Sea $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$. Entonces, $\sum_{i=1}^k X_i = n$.

Ejemplo 5.5 (destreza)

Elegimos n personas al azar de la población, y les preguntamos con qué mano son más hábiles, por ejemplo, específicamente con cuál manejan el cuchillo cuando comen. En este caso, hay 3 resultados posibles: zurdos, derechos o ambidiestros, con probabilidades respectivas 0.10, 0.89 y 0.01. Definimos en este caso 3 variables aleatorias.

Distribución Multinomial

Espacio muestral: Si numeramos los k resultados posibles,

$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : 1 \leq \omega_j \leq k, j = 1, \dots, n\}$ donde ω_j representa el resultado obtenido en la j -ésima repetición del experimento. La probabilidad asociada a cada tira de resultados es

$$P(\omega) = p_{\omega_1} \cdots p_{\omega_n} = p_1^{X_1(\omega)} \cdots p_k^{X_k(\omega)}. \quad (2)$$

Calculamos la función de probabilidad puntual conjunta. Observemos que el rango de $\tilde{\mathbf{X}}$ es

$$R_{\tilde{\mathbf{X}}} = \left\{ \tilde{\mathbf{x}} = (x_1, \dots, x_k) : 0 \leq x_i \leq n, \sum_{i=1}^n x_i = n \right\}$$

(puntos del simplex k dimensional con coordenadas enteras)

Para un $\tilde{\mathbf{x}} \in R_{\tilde{\mathbf{X}}}$ las tiras que pertenecen al evento

$\{\omega \in \Omega : (X_1(\omega), \dots, X_k(\omega)) = (x_1, \dots, x_k)\}$ tienen igual probabilidad dada por (2).

Distribución Multinomial: probabilidad conjunta

Finalmente, la probabilidad conjunta está dada por

$$p_{\tilde{\mathbf{X}}} (x_1 \dots, x_k) = \binom{n}{x_1 \dots x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} = \frac{n!}{x_1! x_2! \cdots x_k!} \prod_{i=1}^k p_i^{x_i},$$

para todo $\tilde{\mathbf{x}} \in R_{\tilde{\mathbf{X}}}$.

En este caso, diremos que el vector $\tilde{\mathbf{X}}$ tiene Distribución Multinomial con parámetros n, p_1, \dots, p_k , y lo notamos $\tilde{\mathbf{X}} \sim \mathcal{M}(n, p_1, \dots, p_k)$,
 $0 < p_i < 1$, $\sum_{i=1}^k p_i = 1$.

Ejemplo 5.5, continuación

Elegimos 10 personas al azar de la población, (con reposición), ¿cuál es la probabilidad de que 9 sean derechos y uno zurdo? En este caso,
 $\tilde{\mathbf{X}} = (X_1, X_2, X_3) \sim \mathcal{M}(n = 10, p_1 = 0.1, p_2 = 0.89, p_3 = 0.01)$ y
 $p_{\tilde{\mathbf{X}}} (1, 9, 0) = \frac{10!}{1! 9! 0!} (0.1)^1 (0.89)^9 (0.01)^0 = 0.3504$

Distribución Multinomial: marginales

La función de probabilidad puntual marginal de cada variable X_i en particular puede obtenerse sumando los valores de la función de probabilidad puntual conjunta sobre todos los otros valores x_j , $j \neq i$. Esta tarea tediosa puede evitarse observando que X_i puede verse como la cantidad de éxitos en n repeticiones independientes del experimento, cada una de las cuales tiene probabilidad p_i de éxito y $1 - p_i$ de fracaso. Por lo tanto, X_i es una variable aleatoria con distribución binomial,

$$X_i \sim \mathcal{B}(n, p_i) \text{ y}$$

$$p_{X_i}(x_i) = \binom{n}{x_i} p_i^{x_i} (1 - p_i)^{n-x_i}, \quad 0 \leq x_i \leq n$$

Ejemplo 5.5, continuación

Elegimos 10 personas al azar de la población, (con reposición), ¿cuál es la probabilidad obtener exactamente 9 derechos? En este caso,

$$X_2 \sim \mathcal{B}(n = 10, p_2 = 0.89) \text{ y } p_{X_2}(9) = \frac{10!}{1!9!} (0.11)^1 (0.89)^9 = 0.3854$$

Vectores Aleatorios (Absolutamente) Continuos

Definición 5.7 (vector aleatorio (absolutamente) continuo)

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. Diremos que el vector aleatorio $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ es **(absolutamente) continuo** si existe una función $f_{\tilde{\mathbf{X}}} : \mathbb{R}^k \rightarrow \mathbb{R}_{\geq 0}$ de forma tal que

$$F_{\tilde{\mathbf{X}}}(x_1, \dots, x_k) = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f_{\tilde{\mathbf{X}}}(t_1, \dots, t_k) dt_1 \cdots dt_k .$$

Notemos que, en tal caso,

$$1 = \lim_{\tilde{x} \rightarrow \infty} F_{\tilde{\mathbf{X}}}(x_1, \dots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\tilde{\mathbf{X}}}(t_1, \dots, t_k) dt_1 \cdots dt_k .$$

Definición 5.8 (función de densidad conjunta)

Toda función $f : \mathbb{R}^k \rightarrow \mathbb{R}_{\geq 0}$ que además integre uno se dice **función de densidad conjunta**:

① $f \geq 0$

② $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, \dots, t_k) dt_1 \cdots dt_k = 1$

Lema 5.6

Si $f_{\tilde{\mathbf{x}}}$ es continua en $\tilde{\mathbf{x}}_0$, entonces

$$\frac{\partial^k}{\partial x_1 \cdots \partial x_k} F_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}_0) = f_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}_0)$$

Lema 5.7

Si $\tilde{\mathbf{X}}$ es un vector aleatorio con función de densidad $f_{\tilde{\mathbf{X}}}$, entonces tenemos que para todo $A \in \mathcal{B}^k$ vale que

$$P(\tilde{\mathbf{X}} \in A) = \int \cdots \int_A f_{\tilde{\mathbf{X}}}(x_1, \dots, x_k) dx_1, \dots dx_k . \quad (3)$$

Dado un vector aleatorio continuo $\tilde{\mathbf{X}}$, su función de densidad conjunta $f_{\tilde{\mathbf{X}}}$ nos permite conocer mediante la relación dada en (1) el valor de las probabilidades $P(X \in A)$ para cualquier boreliano A de \mathbb{R}^k . Aclaremos que sólo estaremos interesados en calcular probabilidades (y por lo tanto integrales) que involucren regiones de tipo I o tipo II, con lo cual las integrales que nos puedan aparecer no serán otras que las que ya saben calcular.

Ejercicio 5.2

Sea (X, Y) un vector aleatorio con función de densidad conjunta dada por

$$f(x, y) = \begin{cases} c(x + y) & \text{si } 0 < x, y < 1 \\ 0 & \text{caso contrario} \end{cases} \quad (4)$$

Hallar c. Calcular (i) $P(X \leq 1/3, Y > 1/2)$, (ii) $P(X + Y < 1)$, (iii) $P(X \leq 1/3)$, (iv) $P(X \leq x)$.

¿Cuánto pueden diferir dos “versiones” de la densidad?

Un vector aleatorio $\tilde{\mathbf{X}} \in \mathbb{R}^k$ es (absolutamente) continuo cuando existe una función boreiana $f : \mathbb{R}^k \rightarrow [0, +\infty)$ que integra uno que denominamos *función de densidad de $\tilde{\mathbf{X}}$* . Para calcular la probabilidad de que $\tilde{\mathbf{X}} \in A$ para todo $A \in \mathcal{B}^k$ tenemos que integrar

$$P(\tilde{\mathbf{X}} \in A) = \int \cdots \int_A f(x_1, \dots, x_k) dx_1, \dots dx_k.$$

Es decir, que todo lo que nos interesa de la función de densidad es cuánto vale su integral en distintos conjuntos boreianos de \mathbb{R}^k .

¿Cuánto pueden diferir entre sí dos funciones $f, g : \mathbb{R}^k \rightarrow [0, +\infty)$ boreianas que son ambas densidades del vector $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$?

Misma respuesta que en el caso de una variable aleatoria (abs) continua: el conjunto

$$E = \left\{ \tilde{\mathbf{x}} \in \mathbb{R}^k : f(\tilde{\mathbf{x}}) \neq g(\tilde{\mathbf{x}}) \right\}$$

es un conjunto de medida cero en \mathbb{R}^k . Y por lo tanto hablaremos de LA función de densidad del vector $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$.

Lema 5.1

Sea $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ un vector aleatorio (absolutamente) continuo con función de densidad boreiana $f : \mathbb{R}^k \rightarrow [0, +\infty)$. La función boreiana $g : \mathbb{R}^k \rightarrow [0, +\infty)$ tal que $\int \cdots \int g(x_1, \dots, x_k) dx_1 \cdots dx_k = 1$ es otra función de densidad para $\tilde{\mathbf{X}}$ si y sólo si

$$E = \left\{ \tilde{\mathbf{x}} \in \mathbb{R}^k : f(\tilde{\mathbf{x}}) \neq g(\tilde{\mathbf{x}}) \right\}$$

es un conjunto de medida cero en \mathbb{R}^k .

Lema 5.2

Sea $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ un vector aleatorio (absolutamente) continuo en \mathbb{R}^k . Entonces, para todo boreliano $E \subset \mathbb{R}^k$ de medida nula vale que

$$P(\tilde{\mathbf{X}} \in E) = 0.$$

Podemos usar este lema para descartar que un vector sea (absolutamente) continuo, como en el Ejemplo 5.6. También podemos usarlo para calcular probabilidades (lo harán en algún ejercicio de la práctica), y también para simplificar el cálculo de densidades conjuntas, como en el ejemplo de cambio de variables no inyectivo (Teorema 5.10).

Repasemos la definición de conjunto de medida cero en \mathbb{R}^k , y veamos qué conjuntos tienen medida cero.

Conjuntos de medida cero en \mathbb{R}^k

Definición: Un conjunto $E \subset \mathbb{R}^k$ se dice de medida nula si dado $\epsilon > 0$ existe una sucesión $(B_n)_{n \in \mathbb{N}}$ de rectángulos (o productos cartesianos) de lados paralelos a los ejes:

$$B_n = I_1^n \times \dots \times I_k^n \text{ donde } I_i^n \text{ es un intervalo en } \mathbb{R}, 1 \leq i \leq k$$

tal que:

$$E \subset \bigcup_{n=1}^{\infty} B_n \quad y \quad \sum_{n=1}^{\infty} |B_n| < \epsilon.$$

donde $|B_n|$ es el "volumen" de B_n dado por

$$|B_n| = \prod_{j=1}^k |I_j^n|$$

si $|I_j^n|$ es la longitud del intervalo I_j^n

Conjuntos de medida cero en \mathbb{R}^k

- Si $E \subset \mathbb{R}^k$ es numerable, entonces tiene medida cero.
- Sea $g : \mathbb{R}^k \rightarrow \mathbb{R}$ continua. Entonces su gráfico

$$G(g) = \left\{ (\tilde{\mathbf{x}}, g(\tilde{\mathbf{x}})) : \tilde{\mathbf{x}} \in \mathbb{R}^k \right\} \subset \mathbb{R}^{k+1}$$

tiene medida cero.

Por lo que una recta en \mathbb{R}^2 , o un plano contenido en \mathbb{R}^3 tienen medida cero.

- También tiene medida cero la imagen de una curva $\gamma : [a, b] \rightarrow \mathbb{R}^2$ de clase C^1 :

$$\{(\gamma_1(t), \gamma_2(t)) : t \in [a, b]\} \subset \mathbb{R}^2$$

Por ejemplo, una circunferencia en \mathbb{R}^2 tiene medida cero.

Relación entre las densidades marginales y conjunta

Observemos que

$$\begin{aligned} P(X \in B) &= P\left((X, Y) \in B \times (-\infty, \infty)\right) \\ &= \int_B \int_{-\infty}^{\infty} f_{XY}(x, y) \, dy \, dx . \\ &= \int_B \underbrace{\left[\int_{-\infty}^{\infty} f_{XY}(x, y) \, dy \right]}_{g(x)} \, dx . \\ &= \int_B g(x) \, dx \end{aligned}$$

Tenemos entonces una fórmula para la función de densidad marginal a partir de la densidad conjunta.

Relación entre las densidades marginales y conjunta

Lema 5.8

Sea (X, Y) un vector aleatorio continuo con función de densidad conjunta dada por f_{XY} , entonces tenemos que las funciones de densidad marginales vienen dadas por

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy ,$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx ,$$

Lo cual es análogo al caso discreto, (X, Y) donde la función de probabilidad puntual marginal vimos que se obtenía a partir de la función de probabilidad puntual conjunta:

$$p_X(x) = \sum_{y \in R_Y} p_{XY}(x, y) , \quad p_Y(y) = \sum_{x \in R_X} p_{XY}(x, y) ,$$

Generalización,

$$f_{X_i}(\textcolor{red}{x}) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{\substack{k-1 \text{ integrales,} \\ \text{salvo la iésima}}} f_{\tilde{\mathbf{X}}}(y_1, \dots, y_{i-1}, \textcolor{red}{x}, y_{i+1}, \dots, y_k) \underbrace{dy_1 \cdots dy_k}_{\text{coord } i \text{ excluida}},$$

En particular, si $\tilde{\mathbf{X}}$ es absolutamente continuo, lo mismo vale para cada una de sus coordenadas X_i , $i = 1, \dots, k$. También lo son todos sus subvectores. También podemos obtener la función de densidad conjunta de subvectores a partir de la $f_{\tilde{\mathbf{X}}}$: y la función de densidad marginal de (X_1, X_3) , por ejemplo, está dada por:

$$f_{X_1, X_3}(x_1, x_3) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{\substack{k-2 \text{ integrales,} \\ \text{salvo la 1 y 3}}} f_{\tilde{\mathbf{X}}}(x_1, \dots, x_k) \underbrace{dx_2 dx_4 \cdots dx_k}_{\text{coord 1 y 3 excluidas}}$$

Definición 5.9 (distribución uniforme)

El vector aleatorio (X, Y) se dice **uniforme en la región** $A \subset \mathbb{R}^2$ si su función de densidad esta dada por

$$f_{XY}(x, y) = \begin{cases} 1/|A| & \text{si } (x, y) \in A \\ 0 & \text{caso contrario,} \end{cases}$$

siendo $|A|$ el área del conjunto A .

Ejercicio 5.3

Sea (X, Y) un vector con distribución uniforme en el círculo de radio uno. Obtenga la función de densidad de cada una de las coordenadas. ¿Resultan uniformes?

Independencia

Proposición 5.6

Sea $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ absolutamente continuo. Son equivalentes

- ① X_1, \dots, X_k son independientes.
- ② la función de densidad conjunta se factoriza
$$f_{X_1, \dots, X_k}(x_1, \dots, x_k) = \prod_{i=1}^k f_{X_i}(x_i)$$
- ③ la función de distribución conjunta se factoriza:
$$F_{X_1, \dots, X_k}(x_1, \dots, x_k) = F_{X_1}(x_1) \cdots F_{X_k}(x_k).$$
- ④ Existen funciones $h_i : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ tales que
$$f_{X_1, \dots, X_k}(x_1, \dots, x_k) = h_1(x_1) \cdots h_k(x_k).$$
- ⑤ Existen funciones $H_i : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ tales que
$$F_{X_1, \dots, X_k}(x_1, \dots, x_k) = H_1(x_1) \cdots H_k(x_k).$$

Observación: El ítem 4 no impone que los factores sean densidades.

3) \Leftrightarrow 5) usando la misma dem. del caso discreto.

1) (\Rightarrow 3) es la **proposición 5.4**

2) \Rightarrow 3)

$$\begin{aligned} F_{\tilde{X}}(x_1, \dots, x_k) &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f_{\tilde{X}}(t_1, \dots, t_k) dt_k \cdots dt_1 \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f_{X_1}(t_1) \cdots f_{X_k}(t_k) dt_k \cdots dt_1 \\ &= \prod_{i=1}^k \int_{-\infty}^{x_i} f_{X_i}(t_i) dt_i = \prod_{i=1}^k F_{X_i}(x_i). \end{aligned}$$

3) \Rightarrow 2) es leer la cuenta ↑ desde el final al principio

4) (\Leftarrow) 2) es la misma cuenta del caso discreto, cambiando sumatorias por integrales.

Proposición 5.7

Sea $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ un vector aleatorio continuo en \mathbb{R}^k con función de densidad conjunta $f_{\tilde{\mathbf{X}}}$. Definimos el soporte de $\tilde{\mathbf{X}}$ como

$$\text{sop}(\tilde{\mathbf{X}}) = \left\{ \tilde{\mathbf{y}} \in \mathbb{R}^k : \int_V f_{\tilde{\mathbf{x}}} (x_1, \dots, x_k) dx_1 \cdots dx_k > 0 \begin{matrix} \text{para todo entorno} \\ V \text{ de } \tilde{\mathbf{y}} \end{matrix} \right\} \quad (5)$$

Entonces, si X_1, \dots, X_k son independientes se tiene que

$\text{sop}(\tilde{\mathbf{X}}) = \text{sop}(X_1) \times \dots \times \text{sop}(X_k)$, donde el soporte de las X_i se define de manera análoga a (5).

Notemos que de esta última proposición se deduce que si $\tilde{\mathbf{X}}$ es un vector aleatorio continuo cuyo soporte no es un **producto cartesiano** (en \mathbb{R}^k) entonces sus coordenadas no son independientes. Esto constituye una manera rápida y eficiente de descartar independencia. No obstante, la recíproca no es cierta. Queda como ejercicio pensar un contraejemplo en donde el soporte de $\tilde{\mathbf{X}}$ sea un **producto cartesiano** pero sus coordenadas no sean independientes.

Para $k=2$ (caso general es análogo).

$$\text{Q.V.Q} \quad \text{sop}(x_1, x_2) = \text{sop}(x_1) \times \text{sop}(x_2) \quad (\text{abierto de } \mathbb{R} \text{ que contiene a } j_1)$$

Sea $(y_1, y_2) \in \text{sop}(x_1, x_2)$. Sean I_1 entorno de y_1
 I_2 entorno de y_2 .

Queremos ver que $P(x_1 \in I_1) > 0$ y $P(x_2 \in I_2) > 0$
ya que eso implica $(y_1, y_2) \in \text{sop}(x_1) \times \text{sop}(x_2)$

Como $I_1 \times I_2$ es un entorno de (y_1, y_2) , por definición
de $\text{sop}(x_1, x_2)$ tenemos:

$$P((x_1, x_2) \in I_1 \times I_2) > 0.$$

Pero $P(x_1 \in I_1, x_2 \in I_2) \stackrel{''}{=} P(x_1 \in I_1)$

Análogamente, $P(x_2 \in I_2) > 0$ ✓

Observemos que
esta tesis
vale siempre.

2 Sea $(y_1, y_2) \in \text{Sop}(x_1) \times \text{Sop}(x_2)$

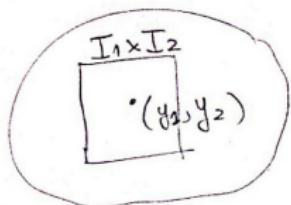
Sea V un entorno de (y_1, y_2) . Queda:

$$P((x_1, x_2) \in V) > 0.$$

Como V es abierto y contiene a (y_1, y_2) existen $\delta_1, \delta_2 > 0$

tales que

$$\underbrace{(y_1 - \delta_1, y_1 + \delta_1)}_{I_1} \times \underbrace{(y_2 - \delta_2, y_2 + \delta_2)}_{I_2} \subset V$$



Como $y_1 \in \text{sop}(x_1)$ e I_1 es un entorno de

y_1 , resulta que $P(x_1 \in I_1) > 0$. Análogamente $P(x_2 \in I_2) > 0$

Luego $P((x_1, x_2) \in V) \geq P((x_1, x_2) \in I_1 \times I_2) = \prod_{i=1}^2 P(x_i \in I_i) > 0$
indep.

Ejercicio 5.4

Sea (X, Y) un vector aleatorio con función de densidad conjunta dada por

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y} & \text{si } 0 < x, y \\ 0 & \text{caso contrario} \end{cases} \quad (6)$$

¿Son X e Y independientes? Calcular (i) $P(X > 1, Y < 1)$, (ii) $P(X \leq Y)$.

Ejercicio 5.5

Sea (X, Y) un vector aleatorio con función de densidad conjunta dada por

$$f(x, y) = \begin{cases} 24xy & \text{si } 0 < x, y < 1, 0 \leq x + y \leq 1 \\ 0 & \text{caso contrario} \end{cases} \quad (7)$$

¿Son X y Y independientes? Obtenga las densidades marginales.

Observación 5.1

Vimos que

- (*caso discreto*) si $\tilde{\mathbf{X}}$ es discreto, lo mismo vale para cada una de sus coordenadas X_i , $i = 1, \dots, k$.
- (*caso continuo*) si $\tilde{\mathbf{X}}$ es absolutamente continuo, lo mismo vale para cada una de sus coordenadas X_i , $i = 1, \dots, k$.

Al revés, tenemos:

- (*caso discreto*) si cada X_i , $i = 1, \dots, k$ es una variable discreta, el vector $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ resulta discreto.
- (*caso continuo*) A diferencia del caso discreto, no vale la recíproca: si cada X_i , $i = 1, \dots, k$ es una variable absolutamente continua, el vector $\tilde{\mathbf{X}} = (X_1, \dots, X_k)$ no es necesariamente absolutamente continuo. Veamos un contraejemplo.

Concatenando variables (abs) continuas no necesariamente conseguimos un vector (absolutamente) continuo

Ejemplo 5.6

Sea $\tilde{\mathbf{X}} = (X, Y)$ con $X \sim \mathcal{U}(0, 1)$, $Y \sim \mathcal{U}(0, 1)$.

- ① Si X es independiente de Y , el vector (X, Y) es absolutamente continuo, según la Proposición 5.6, más aún su densidad conjunta es $f_{XY}(x, y) = I_{(0,1)}(x)I_{(0,1)}(y)$.
- ② Tomemos $Y = X$. En este caso el vector (X, Y) está soportado en la diagonal $\Delta = \{(x, y) \in \mathbb{R}^2 : x = y\}$ que tiene área cero, y por lo tanto, no es absolutamente continuo.

Funciones de vectores aleatorios

Definición 5.10 (Funciones Boreelianas a valores vectoriales)

Diremos que una función $g : \mathbb{R}^k \rightarrow \mathbb{R}^j$ es *boreiana* (o medible Borel) si para todo $A \in \mathcal{B}(\mathbb{R}^j)$ vale que $g^{-1}(A) \in \mathcal{B}(\mathbb{R}^k)$.

Recordemos el

Lema 5.4

Sea $g : \mathbb{R}^k \rightarrow \mathbb{R}^j$ una función Boreiana. Sea $\tilde{\mathbf{X}}$ un vector aleatorio k -dimensional, entonces $\tilde{\mathbf{Y}} = g(\tilde{\mathbf{X}})$ es un vector aleatorio j dimensional.

Funciones de vectores aleatorios

Caso discreto: Si $\tilde{\mathbf{X}}$ es un vector aleatorio discreto, entonces $\tilde{\mathbf{Y}} = g(\tilde{\mathbf{X}})$ es necesariamente un vector aleatorio discreto, $R_{\tilde{\mathbf{Y}}} = \{g(\mathbf{k}_i) : \mathbf{k}_i \in R_{\tilde{\mathbf{X}}}\}$ y

$$p_{\tilde{\mathbf{Y}}}(\mathbf{y}) = \sum_{\substack{\mathbf{k}_i \in R_{\tilde{\mathbf{X}}}: \\ g(\mathbf{k}_i) = \mathbf{y}}} p_{\tilde{\mathbf{X}}}(\mathbf{k}_i)$$

Caso absolutamente continuo: Si $\tilde{\mathbf{X}}$ es un vector aleatorio (absolutamente) continuo, entonces $\tilde{\mathbf{Y}} = g(\tilde{\mathbf{X}})$ no necesariamente es un vector aleatorio (absolutamente) continuo. Si $g : \mathbb{R}^k \rightarrow \mathbb{R}^j$ toma finitos o numerables valores, $\tilde{\mathbf{Y}}$ será discreto. Por ejemplo, si $g : \mathbb{R}^k \rightarrow \mathbb{R}$, $g(\mathbf{x}) = c$ para todo \mathbf{x} , o sea, si g es una función constante, entonces $\mathbf{Y} = g(\tilde{\mathbf{X}})$ es discreta con rango $R_Y = \{c\}$.

Teorema de cambio de variable

Teorema 5.9 (Cambio de Variable)

Sea $\tilde{\mathbf{X}}$ un vector aleatorio (absolutamente) continuo en \mathbb{R}^k .

Sea $G_1 \subset \mathbb{R}^k$ abierto tal que $P(\tilde{\mathbf{X}} \in G_1) = 1$.

Sea $g : G_1 \rightarrow G_2 \subset \mathbb{R}^k$ función biyectiva de clase C^1 con G_2 abierto,
 $\det[Dg(\tilde{\mathbf{x}})] \neq 0, \forall \tilde{\mathbf{x}} \in G_1$.

Sea $g^{-1} : G_2 \rightarrow G_1$ su inversa. Entonces $\tilde{\mathbf{Y}} = g(\tilde{\mathbf{X}})$ es un vector aleatorio
absolutamente continuo con función de densidad

$$f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}) = f_{\tilde{\mathbf{X}}}(g^{-1}(\tilde{\mathbf{y}})) \cdot |\det[Dg^{-1}(\tilde{\mathbf{y}})]| \cdot I_{G_2}(\tilde{\mathbf{y}}).$$

Cambio de Variable, caso no inyectivo

Teorema 5.10 (Cambio de Variable, caso no inyectivo)

Sea $\tilde{\mathbf{X}}$ un vector aleatorio (absolutamente) continuo en \mathbb{R}^k .

Sean $U_1, \dots, U_n \subset \mathbb{R}^k$ abiertos disjuntos tales que $P(\tilde{\mathbf{X}} \in \bigcup_{j=1}^n U_j) = 1$.

Sea $g : \bigcup_{j=1}^n U_j \rightarrow \mathbb{R}^k$ tal que si llamamos g_j a la restricción de g a U_j , o sea $g_j : U_j \rightarrow \mathbb{R}^k$ con $g_j(\tilde{\mathbf{x}}) = g(\tilde{\mathbf{x}})$, $\forall \tilde{\mathbf{x}} \in U_j$ es una función inyectiva de clase C^1 con $\det[Dg_j(\tilde{\mathbf{x}})] \neq 0$, $\forall \tilde{\mathbf{x}} \in U_j$.

Sea $V_j = g_j(U_j)$ abierto de modo que $g_j : U_j \rightarrow V_j$ sea biyectiva, y sea $g_j^{-1} : V_j \rightarrow U_j$ su inversa. Entonces $\tilde{\mathbf{Y}} = g(\tilde{\mathbf{X}})$ es un vector aleatorio absolutamente continuo con función de densidad

$$f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}) = \sum_{j=1}^n f_{\tilde{\mathbf{X}}} \left(g_j^{-1}(\tilde{\mathbf{y}}) \right) \cdot \left| \det \left[Dg_j^{-1}(\tilde{\mathbf{y}}) \right] \right| \cdot I_{V_j}(\tilde{\mathbf{y}}).$$

Suma de variables aleatorias: caso discreto

Sean X e Y variables aleatorias discretas, definimos $Z = X + Y$. Entonces Z es discreta con $R_Z \subset \{x + y : x \in R_X, y \in R_Y\}$ y

$$\begin{aligned} p_Z(z) &= P(X + Y = z) = \sum_{\substack{x \in R_X, y \in R_Y: \\ x+y=z}} P(X = x, Y = y) \\ &= \sum_{x \in R_X} P(X = x, Y = z - x) = \sum_{x \in R_X} p_{XY}(x, z - x) \end{aligned}$$

Además, si X e Y son independientes,

$$p_Z(z) = \sum_{x \in R_X} p_X(x) p_Y(z - x)$$

Suma de variables aleatorias: caso absolutamente continuo

Proposición 5.8

Sea (X, Y) un vector aleatorio continuo, con densidad f_{XY} . Tenemos entonces que $Z = X + Y$ es una variable aleatoria continua con densidad dada por

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{XY}(x, z - x) dx .$$

Además, si X e Y son independientes,

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z - x) dx = \int_{-\infty}^{+\infty} f_X(z - y)f_Y(y) dy .$$

Demostración.

Sea $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ dada por $g(x, y) = (x + y, y)$ con inversa $g^{-1}(z, w) = (z - w, w)$. Como $Dg^{-1}(z, w) = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$, por el Teorema de cambio de variables tenemos,

$$f_{ZW}(z, w) = f_{XY}(g^{-1}(z, w)) \cdot |\det [Dg^{-1}(z, w)]| = f_{XY}(z - w, w)$$

y la marginal

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{ZW}(z, w) dw = \int_{-\infty}^{+\infty} f_{XY}(z - w, w) dw$$

La otra igualdad se obtiene a partir de esta por cambio de variables (unidimensional). □

Definición 5.11 (convolución)

Dadas $f, g : \mathbb{R} \rightarrow \mathbb{R}$ integrables, se llama la **convolución** de f y g , y se denota por $f * g$ a

$$f * g(x) = \int_{-\infty}^{+\infty} f(x - y) g(y) dy$$

Luego, probamos en la Proposición 5.8 que si X e Y son absolutamente continuas e independientes, la densidad de $Z = X + Y$ está dada por la convolución de sus densidades, es decir,

$$f_Z = f_X * f_Y$$

Ejercicio 5.6

Sean $X \sim \Gamma(\alpha, \lambda)$ $Y \sim \Gamma(\beta, \lambda)$, variables aleatorias independientes. ¿Qué distribución tiene $Z = X + Y$? ¿Pertenece a una familia conocida? Generalizar.

Ejercicio 5.7

Sea (X, Y) un vector aleatorio absolutamente continuo con función de densidad

$$f_{XY}(x, y) = \begin{cases} \frac{1}{2x^2} & \text{si } |x| < 1, 0 < y < x^2 \\ 0 & \text{en caso contrario} \end{cases}$$

- ① Hallar f_X y f_Y . ¿Son X e Y independientes?
- ② Probar que $\frac{Y}{X^2}$ tiene distribución $\mathcal{U}[0, 1]$ y es independiente de X .
- ③ Hallar la densidad conjunta del vector $(X^2, \frac{Y}{X^2})$.

Ejemplo de cambio de variables (no inyectivo)

Ejemplo: Sean X e Y variables aleatorias independientes, $X \sim N(0,1)$, $Y \sim N(0,1)$. Nos interesa estudiar las siguientes variables aleatorias:

$$Z = X^2 + Y^2 \quad \text{y} \quad W = \frac{X}{Y}. \quad \text{Queremos hallar}$$

f_{ZW} , y luego $f_Z f_W$.

$$f_{XY}(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \quad \text{porque } X \text{ e } Y \text{ son independientes}$$

Sea $g(x,y) = (X^2 + Y^2, \frac{X}{Y}) = (Z, W)$. Esta función está definida siempre que Y no sea nula.

$$g: \mathbb{R} \times (-\infty, 0) \cup (0, +\infty) \rightarrow \mathbb{R}^2$$

Sean $U_1 = \mathbb{R} \times (-\infty, 0)$
 $U_2 = \mathbb{R} \times (0, +\infty)$

Ejemplo de cambio de variables (no inyectivo, cont.)

$g: \underbrace{\mathbb{R} \times (-\infty, 0)}_{U_1} \cup \underbrace{\mathbb{R} \times (0, +\infty)}_{U_2} \rightarrow \mathbb{R}^2$ está bien definida (A).
 $g(x, y) = (x^2 + y^2, \frac{x}{y}) = (z, w)$

Observemos que $\#((x, y) \in U_1 \cup U_2) = 1$.

Hallemos la inversa
 $w = \frac{x}{y} \Leftrightarrow x = wy$ reemplazamos en $z = x^2 + y^2 = w^2y^2 + y^2$

$$\Rightarrow z = y^2(1 + w^2) \Rightarrow y^2 = \frac{z}{1 + w^2}, \text{ lo cual está bien definido si } z > 0$$

Esto restringe el codominio de g . Luego $|y| = \sqrt{\frac{z}{1+w^2}}$. Hay 2 inversas,

dependiendo del signo de y . Luego (A) no es inyectiva. Son biyectivas las 2 restricciones. Justamente U_1 y U_2 son dos abiertos sobre los que g es biyectiva.

Ejemplo de cambio de variables (no inyectivo, cont.)

Sean $g_1 := g|_{U_1} : \underbrace{\mathbb{R} \times (-\infty, 0)}_{U_1} \rightarrow (0, +\infty) \times \mathbb{R}$

$$g_1(x, y) = (x^2 + y^2, \frac{x}{y}) = (z, w) \quad (\text{B})$$

y su inversa

$$g_1^{-1} : (0, +\infty) \times \mathbb{R} \rightarrow U_1 = \mathbb{R} \times (-\infty, 0)$$

$$g_1^{-1}(z, w) = \left(-\frac{w\sqrt{z}}{\sqrt{1+w^2}}, \frac{\sqrt{z}}{\sqrt{1+w^2}} \right) = (-w, 1) \cdot \frac{\sqrt{z}}{\sqrt{1+w^2}}$$

$$g_2 := g|_{U_2} : \underbrace{\mathbb{R} \times (0, +\infty)}_{U_2} \rightarrow (0, +\infty) \times \mathbb{R} \text{ con la misma fórmula} \quad (\text{B})$$

$$g_2^{-1} : (0, +\infty) \times \mathbb{R} \rightarrow U_2 = \mathbb{R} \times (0, +\infty)$$

$$g_2^{-1}(z, w) = \left(\frac{w\sqrt{z}}{\sqrt{1+w^2}}, \frac{\sqrt{z}}{\sqrt{1+w^2}} \right) = (w, 1) \frac{\sqrt{z}}{\sqrt{1+w^2}}$$

Vimos

$$\begin{aligned} |y| &= \frac{\sqrt{z}}{\sqrt{1+w^2}} \\ x &= wy. \end{aligned}$$

Ejemplo de cambio de variables (no inyectivo, cont.)

Nota: Observemos que hemos excluido el conjunto $\{z=0\} = \{0\} \times \mathbb{R}$ del codominio de g . Lo hicimos para que $g_1(V_1) = g_2(V_2) = V_1 = V_2 = (0, +\infty) \times \mathbb{R}$ resulte abierto y estemos en las hipótesis del Teorema de Cambio de Variables, caso no inyectivo (Teo 5.10). Lo cual no es un problema pues $P((z, w) \in V_1) = 1$ y sabemos que la densidad conjunta f_{zw} queda determinada salvo en conjuntos de medida cero ($\{0\} \times \mathbb{R}$ es un conjunto de medida cero).

El TCV no inyectivo nos permite asegurar que:

$$(c) f_{zw}(z, w) = \sum_{j=1}^2 f_{xy}(g_j^{-1}(z, w)) |Jg_j^{-1}(z, w)| \cdot I_{V_j}(z, w). \quad \text{Calculemos}$$

Ejemplo de cambio de variables (no inyectivo, cont.)

$$g_1^{-1}(z,w) = (-w, -1) \frac{\sqrt{z}}{\sqrt{w^2+1}}$$

$$f_{xy}(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}$$

Si sumamos sus cuadrados, obtenemos $(w^2+1)\frac{z}{w^2+1} = z$. Lo mismo para

$$g_2^{-1}(z,w) = (w, 1) \frac{\sqrt{z}}{\sqrt{w^2+1}}$$

Como calcular $|Jg_1^{-1}(z,w)|$ es un hio, podemos usar $|Jg_1^{-1}(z,w)| = \frac{1}{|Jg_1(g_1(z,w))|}$

$$g_1(x,y) = (x^2+y^2, xy)$$

$$Dg_1(x,y) = \begin{bmatrix} 2x & 2y \\ y & -x/y^2 \end{bmatrix} = Dg_2(x,y) \Rightarrow |\det(Dg_1(x,y))| = \left| -\frac{2x^2}{y^2} - 2 \right| = 2\left(\frac{x^2}{y^2} + 1\right)$$

$$|Jg_1(g_1(z,w))| = 2(w^2+1)$$

Ejemplo de cambio de variables (no inyectivo, cont.)

Entradas

$$|\operatorname{J}g_1^{-1}(z,w)| = \frac{1}{2(w^2+1)} = |\operatorname{J}g_2^{-1}(z,w)|. \quad \text{Reemplazamos todo en la fórmula (c)}$$

$$\begin{aligned} f_{zw}(z,w) &= \sum_{j=1}^2 f_{xy}(g_j^{-1}(z,w)) |\operatorname{J}g_j^{-1}(z,w)| I_{V_i}(z,w) \\ &= \sum_{j=1}^2 \frac{1}{2\pi} e^{-z/2} \cdot \frac{1}{2(w^2+1)} I_{(0,+\infty)}(z) I_R(w) \\ &= \frac{2}{2\pi} e^{-z/2} \cdot \frac{1}{2(w^2+1)} I_{(0,+\infty)}(z) = \underbrace{\frac{1}{2} e^{-z/2} I_{(0,+\infty)}(z)}_{f_z(z)} \cdot \underbrace{\frac{1}{\pi(w^2+1)}}_{f_w(z)} \end{aligned}$$

Ejemplo de cambio de variables (no inyectivo, cont.)

Luego $Z \sim \text{Exp}(1/2)$ $W \sim \text{Cauchy}(0,1)$ y son independientes.

Distribución Normal Bivariada

Un vector aleatorio (X, Y) tiene distribución normal bivariada si su función de densidad conjunta es

$$f_{XY}(x, y) = \frac{1}{C} e^{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) \right]}.$$

con $C = 2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}$. La función de densidad depende de cinco

constantes:

$$\begin{aligned} -\infty < \mu_X < \infty &\quad -\infty < \mu_Y < \infty \\ \sigma_X > 0 &\quad \sigma_Y > 0 \quad -1 < \rho < 1 \end{aligned}$$

Las curvas de nivel vienen dadas por

$$\left(\frac{x-\mu_X}{\sigma_X} \right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) = \text{constante}$$

Las curvas de nivel vienen dadas por

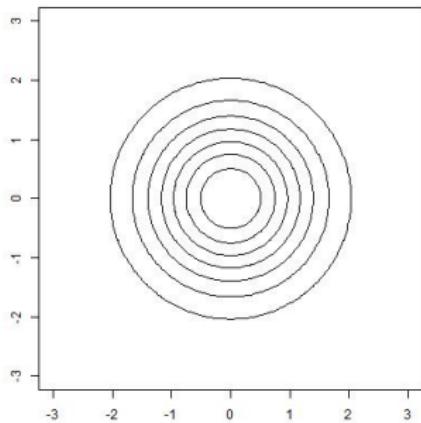
$$\left(\frac{x - \mu_X}{\sigma_X}\right)^2 + \left(\frac{y - \mu_Y}{\sigma_Y}\right)^2 - 2\rho \left(\frac{x - \mu_X}{\sigma_X}\right) \left(\frac{y - \mu_Y}{\sigma_Y}\right) = \text{constante}$$

Son elipses, con centro en (μ_X, μ_Y) . Si $\rho = 0$, los ejes de la elipse son paralelos a los ejes cartesianos x e y , y si $\rho \neq 0$, están inclinados.

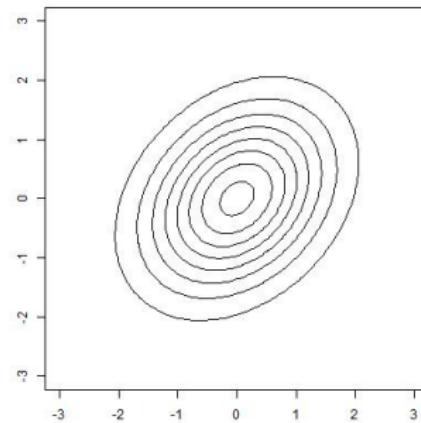
Miremos los gráficos de las curvas de nivel para unas cuantas combinaciones de valores de los parámetros.

Fijamos $(\mu_X, \mu_Y) = (0, 0)$ y $\sigma_X = \sigma_Y = 1$, variaremos ρ .

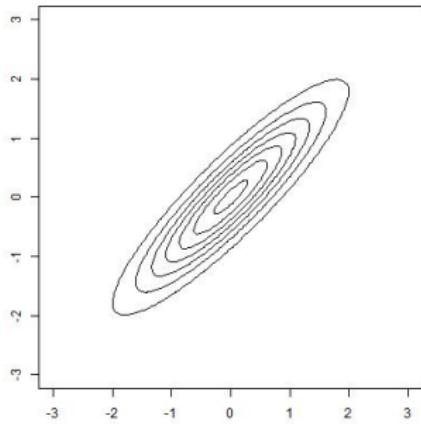
- a) $\rho = 0$
- b) $\rho = 0.3$
- c) $\rho = 0.9$
- d) $\rho = -0.6$



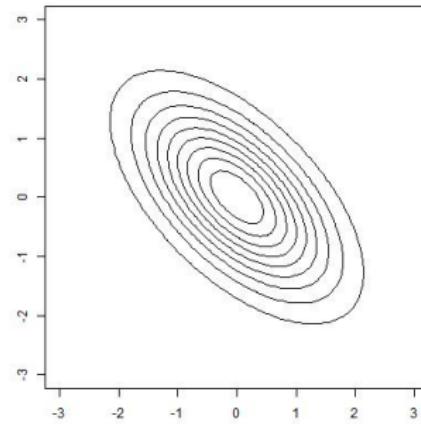
a) $\rho = 0$



b) $\rho = 0.3$



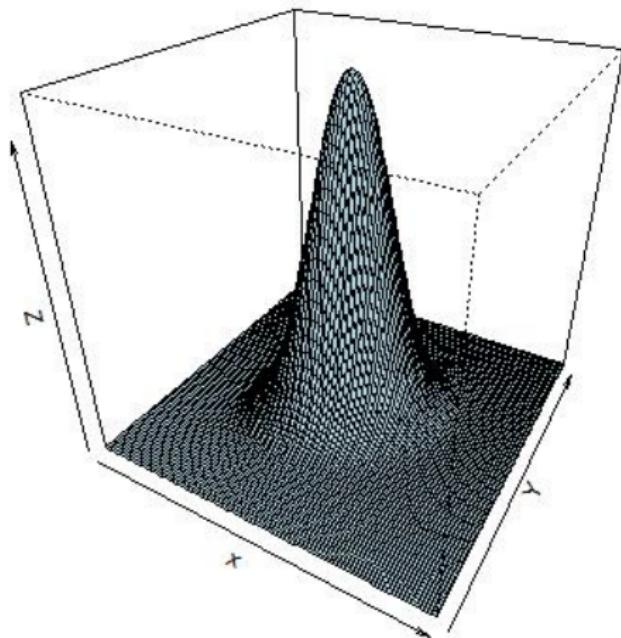
c) $\rho = 0.9$



d) $\rho = -0.6$

Densidad normal bivariada

$\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, $\rho = 0, 3$.



Las distribuciones marginales de X e Y son $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, respectivamente. Veámoslo. La densidad marginal de X es

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

Para facilitar la notación, llamamos $u = (x - \mu_X)/\sigma_X$ y hacemos el cambio de variables en la integral: $v = (y - \mu_Y)/\sigma_Y$ con $dv = \frac{1}{\sigma_Y} dy$

$$f_X(x) = \frac{1}{2\pi\sigma_X\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}(u^2 + v^2 - 2\rho uv)\right] dv$$

Para calcular la integral, completamos cuadrados en v . Usamos la igualdad:

$$u^2 + v^2 - 2\rho uv = (v - \rho u)^2 + u^2(1 - \rho^2)$$

tenemos

$$f_X(x) = \frac{1}{2\pi\sigma_X\sqrt{1-\rho^2}} e^{-u^2/2} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}(v - \rho u)^2\right] dv$$

Finalmente, podemos reconocer la integral que nos queda como la integral de una densidad normal, con parámetros $\mu = \rho u$ y $\sigma^2 = (1 - \rho^2)$.

Copiamos

$$f_X(x) = \frac{1}{2\pi\sigma_X\sqrt{1-\rho^2}} e^{-u^2/2} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}(v-\rho u)^2\right] dv$$

Finalmente, podemos reconocer la integral que nos queda como la integral de una densidad normal con parámetros $\mu = \rho u$ y $\sigma^2 = (1 - \rho^2)$.

Arreglamos la constante que le falta para que sea la densidad normal, y usamos que ¡integra uno!

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-u^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(v-\rho u)^2\right] dv$$

Finalmente reemplazamos de vuelta: $u = (x - \mu_X)/\sigma_X$ y tenemos

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} e^{-(1/2)[(x-\mu_X)^2/\sigma_X^2]}$$

O sea, $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$.

Marginales y conjunta

Obtuvimos $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$. La marginal de Y (es simétrico el problema), también es normal $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Observemos que no importa el valor de ρ para el cálculo de las marginales. Por ejemplo, en los 4 casos en los que vimos las curvas de nivel para las normales bivariadas, obtendríamos que ambas marginales son $\mathcal{N}(0, 1)$.

Independencia en la normal bivariada

(X, Y) tienen distribución normal bivariada. ¿Cuándo son X e Y independientes? Lo serán si

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

$$f_{XY}(x, y) = \frac{1}{C} e^{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X}\right) \left(\frac{y-\mu_Y}{\sigma_Y}\right) \right]}.$$

con $C = 2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}$.

$$f_X(x)f_Y(y) = \frac{1}{\sigma_X\sqrt{2\pi}} e^{-\frac{1}{2} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 \right]} \cdot \frac{1}{\sigma_Y\sqrt{2\pi}} e^{-\frac{1}{2} \left[\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \right]}.$$

Son independientes si y sólo si $\rho = 0$.

Notación

Notación: $(X, Y) \sim \mathcal{N}_2 \left((\mu_X, \mu_Y), \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right)$.

En particular, si X e Y son normales estándares independientes, tenemos que $(X, Y) \sim \mathcal{N}_2 ((0, 0), I)$.

Si llamamos $\Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$, vemos que Σ resulta simétrica y definida positiva (es decir, sus autovalores son mayores que cero).

Además puede verse que la expresión que aparece en el exponente de la exponencial en la densidad conjunta del vector normal bivariado puede escribirse matricialmente así:

$$-\frac{1}{2} [(x, y) - (\mu_X, \mu_Y)] \cdot \Sigma^{-1} \cdot \begin{bmatrix} x - \mu_X \\ y - \mu_Y \end{bmatrix},$$

$$\text{y } \det(\Sigma) = \sigma_X^2 \sigma_Y^2 (1 - \rho^2)$$

Distribución Normal Multivariada

Definición 5.12 (distribución normal multivariada)

Sea $\Sigma \in \mathbb{R}^{k \times k}$ una matriz simétrica definida positiva y $\mu \in \mathbb{R}^k$ y sea $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)$ un vector aleatorio en \mathbb{R}^k con densidad conjunta dada por

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{k/2} [\det(\Sigma)]^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mu)^t \Sigma^{-1} (\mathbf{y} - \mu) \right\}.$$

Entonces decimos que el vector \mathbf{Y} tiene **distribución normal multivariada k -dimensional** (bivariada si $k = 2$) de parámetros μ y Σ y lo denotamos por $\mathbf{Y} \sim \mathcal{N}_k(\mu, \Sigma)$.

Otros resultados importantes de la normal

- Si $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ son independientes entonces $aX + bY + c \sim \mathcal{N}(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2)$ para todo $a, b, c \in \mathbb{R}$ con $a \neq 0$ ó $b \neq 0$ (ejercicio de la práctica).
- Sean $U \sim \mathcal{N}(0, 1)$ y $V \sim \chi_n^2 = \Gamma(n/2, 1/2)$ con U y V independientes. Se define la **distribución t de Student con n grados de libertad**, que simbolizaremos con t_n , como la distribución de

$$T = \frac{U}{\sqrt{V/n}}$$

La densidad de T es

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2}) \sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

El gráfico de esta densidad es simétrico respecto al origen (función par) y con forma de campana. Se puede probar que cuando $n \rightarrow +\infty$, la densidad t_n converge puntualmente a la densidad normal estándar.

Estadísticos de Orden

Nos interesa ahora otro grupo de transformaciones de las variables aleatorias continuas X_1, X_2, \dots, X_n . Asumimos que son independientes con la misma función de distribución F . Se definen los **estadísticos de orden** $X^{(1)}, \dots, X^{(n)}$ como aquellas variables aleatorias que se obtienen ordenando las X_i de manera creciente. En particular, tenemos que

$$X^{(1)} = \min_{1 \leq i \leq n} X_i \quad (8)$$

$$X^{(n)} = \max_{1 \leq i \leq n} X_i \quad (9)$$

y $X^{(1)} \leq \dots \leq X^{(n)}$. Es interesante notar que las funciones de distribución de estas variables aleatorias no son las mismas que las originales.

Ejemplo 5.7

Paulina es la encargada de un bar que tiene n mesas. Ella cierra el bar y puede irse a su casa (que es lo que más desea a las 12 de la noche) cuando se va la última mesa. Como el bar es muy exitoso, las n mesas están llenas a las 12 de la noche cuando ella cierra la puerta y no deja entrar a nuevos comensales. Sea X_i el horario, a partir de las 12, en el que se retira el último cliente de la mesa i -ésima. Podemos asumir que estas variables aleatorias son independientes e idénticamente distribuidas, con distribución F conocida. El horario en el que Paulina cierra el local, sin embargo, tendrá la distribución del máximo, $X^{(n)}$, que, lamentablemente para ella, no coincide con F .

La función de distribución se puede calcular, observando que $X^{(n)} \leq u$ si y sólo si $X_i \leq u$ para todo i . Se puede probar (es un ejercicio de la práctica) que

$$F_{X^{(k)}}(x) = P(X^{(k)} \leq x) = \sum_{j=k}^n \binom{n}{j} F(x)^j (1 - F(x))^{n-j}$$

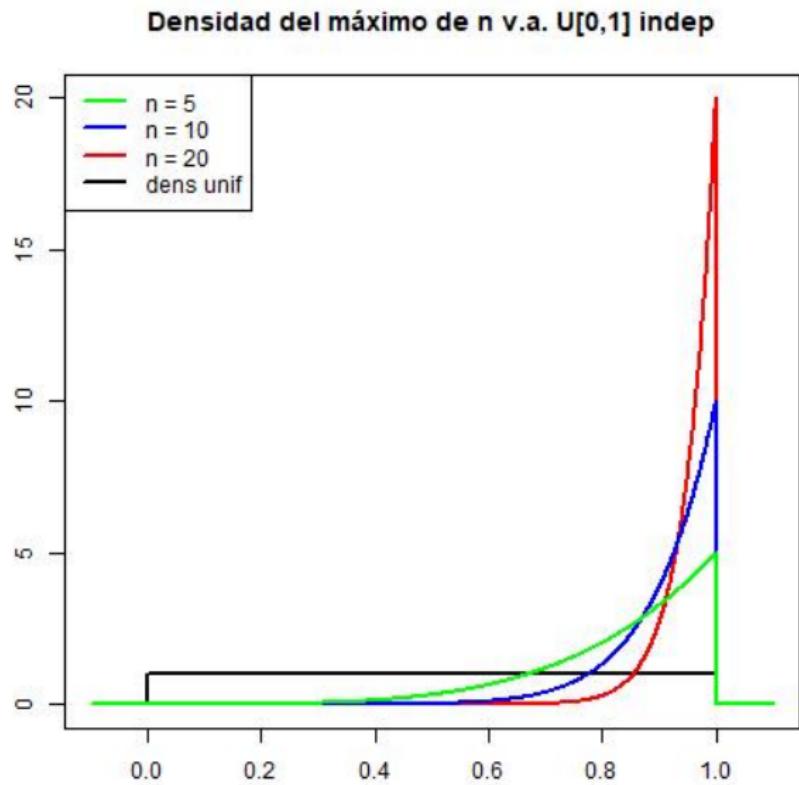
y derivando se calcula la densidad.

Caso uniformes

Es un ejercicio de la Práctica 5, probar que si las X_i tienen distribución uniforme en el intervalo $[0,1]$ entonces para cada $k = 1, \dots, n$ la variable aleatoria $X^{(k)}$ tiene distribución $\beta(k, n - k + 1)$.

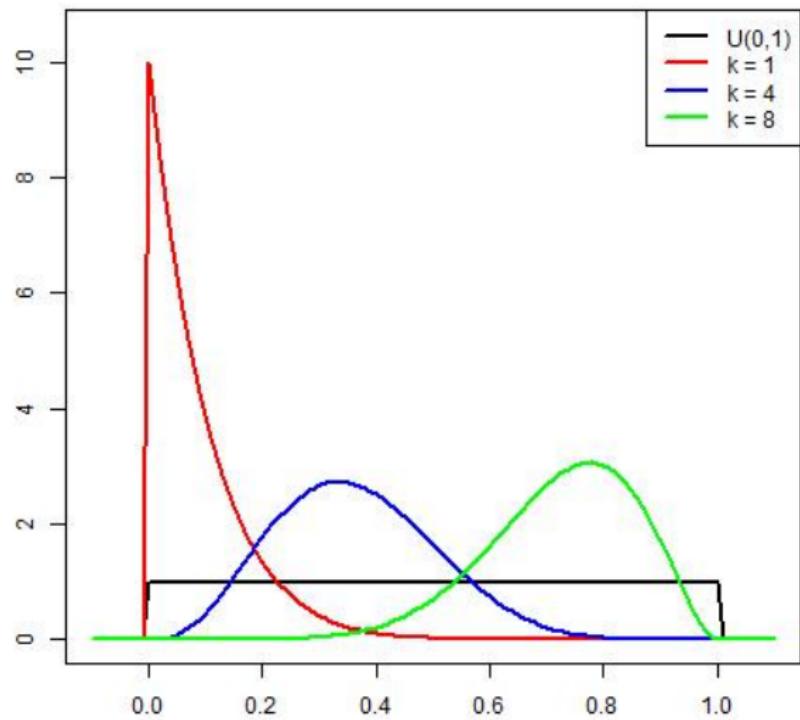
Para ejemplificar, graficamos la densidad del máximo para varios valores de n : 5, 10 y 15. Y luego, graficamos la densidad del k -ésimo estadístico de orden para $n = 10$ y $k = 1, 4, 8$.

Densidad del máximo $X^{(n)}$ de n v.a. independientes $\mathcal{U}(0, 1)$



Densidad de $X^{(k)}$, de n v.a. indep $\mathcal{U}(0, 1)$

$n = 10$



A esta altura, en la que hablamos bastante de la distribución del máximo y mínimo, suele surgir una pregunta que es bastante razonable.



Mmmmm... no entiendo.

Si el mínimo de n variables aleatorias **ES** alguna de esas variables aleatorias, y todas tienen la misma distribución F , ¿cómo es que la distribución del mínimo no es también esa F ?

Para responderla, miremos el siguiente ejemplo con cuidado.

Ejemplo 5.8 (María y sus hijos bebés)

María tiene tres hijos bebés: Ana, Braulio y Cecilia. Los acuesta a dormir de noche. Como canta muy lindo, los tres se duermen apenas ella termina la canción. La cantidad de horas que duerme de corrido cada bebé es una variable aleatoria discreta que toma los valores 6 y 8 con probabilidad $1/2$ cada uno. Asumimos que estas tres variables aleatorias son independientes entre sí.

- Hallar la distribución de la cantidad de horas que duerme de corrido la madre, si se duerme en el mismo momento que sus tres hijos (está muy cansada) y se despierta cuando se despierta el primero de sus tres hijos.
- Comparar la probabilidad de que Ana duerma 8 horas corridas, con la probabilidad de que la madre duerma 8 horas corridas.

a) Numeramos a los bebés 1,2,3. Sean

X_i = cantidad de horas de sueño que duerme de corrido el i -ésimo bebé

$$i=1,2,3$$

X_1, X_2, X_3 son va indep con función de probabilidad puntual:

k	$p_{X_i}(k)$
6	$\frac{1}{2}$
8	$\frac{1}{2}$

$$R_y = \{6, 8\}$$

Sea

y = cantidad de horas de sueño corridas de María

$$y = \min \{X_1, X_2, X_3\}$$

$$p_y(8) = P(\min\{X_1, X_2, X_3\} = 8) = P\left(\bigcap_{i=1}^3 \{X_i = 8\}\right)$$

$$= \prod_{i=1}^3 p_{X_i}(8) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

por independencia

$$p_y(6) = 1 - p_y(8) = \frac{7}{8}. \text{ Luego:}$$

k	$p_y(k)$
6	$7/8$
8	$1/8$

Vemos que la distribución de X_1 es distinta de la de y .

$$p_y(8) = P(\min\{X_1, X_2, X_3\} = 8) = P\left(\bigcap_{i=1}^3 \{X_i = 8\}\right)$$

$$= \prod_{i=1}^3 p_{X_i}(8) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

por independencia

$$p_y(6) = 1 - p_y(8) = \frac{7}{8}. \text{ Luego:}$$

k	$p_y(k)$	Vemos que la distribución de X_1 es distinta de la de y .
6	$7/8$	
8	$1/8$	b) $P(\text{María duerme } 8 \text{ h.}) = p_y(8) = \frac{1}{8}$

$P(\text{Ana duerme } 8 \text{ h.}) = p_{X_1}(8) = \frac{1}{2}$

En este ejemplo vemos que lo que cambia la distribución del mínimo es que, por supuesto, el mínimo es **SISTEMÁTICAMENTE** menor que el resto de las variables aleatorias. Por eso, si comparamos la cantidad de horas que duerme María vemos que la probabilidad de que ella duerma 8 horas corridas es mucho menor que la de que Ana lo haga. Lo mismo pasa en el resto de los ejemplos que vimos, lo que pasa es que este es muy simple, entonces se puede ver lo que sucede. De hecho, podemos hacer la siguiente construcción.

Ejercicio 5.8 (Acoplamiento)

Hagamos la siguiente construcción de las variables aleatorias a partir de una variable $U \sim U(0, 1)$. Sean

$$X_1 = 6 \cdot I_{(0, \frac{1}{2})}(U) + 8 \cdot I_{[\frac{1}{2}, 1]}(U)$$

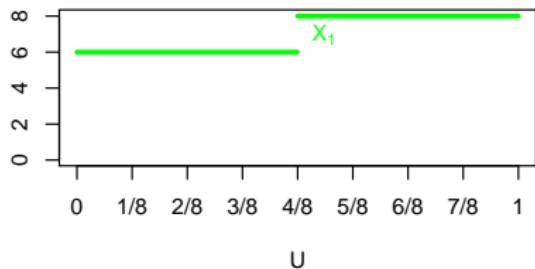
$$X_2 = 6 \cdot I_{(0, \frac{1}{4}) \cup [\frac{1}{2}, \frac{3}{4}]}(U) + 8 \cdot I_{[\frac{1}{4}, \frac{1}{2}) \cup [\frac{3}{4}, 1]}(U)$$

$$X_3 = 6 \cdot I_{(0, \frac{1}{8}) \cup [\frac{2}{8}, \frac{3}{8}) \cup [\frac{4}{8}, \frac{5}{8}) \cup [\frac{6}{8}, \frac{7}{8})}(U) + 8 \cdot I_{[\frac{1}{8}, \frac{2}{8}) \cup [\frac{3}{8}, \frac{4}{8}) \cup [\frac{5}{8}, \frac{6}{8}) \cup [\frac{7}{8}, 1]}(U)$$

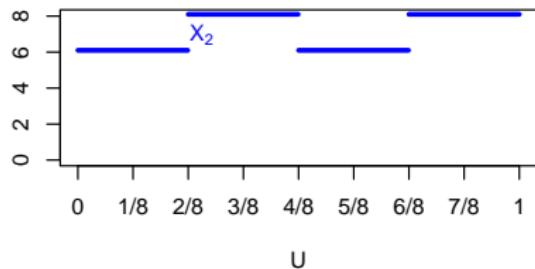
- (a) Verificar que X_1, X_2, X_3 son v.a. que tienen la misma distribución que las horas de sueño del Ejemplo 5.8.
- (b) Probar que son v.a. independientes. (Sugerencia: Tratar de hacer la menor cantidad de cuentas posibles para demostrar que son independientes)
- (c) Graficarlas en un mismo eje de coordenadas, y graficar el mínimo de ellas, $Y = \min\{X_1, X_2, X_3\}$, la cantidad de horas de sueño corridas de María.

Graficamos las variables aleatorias

X_1

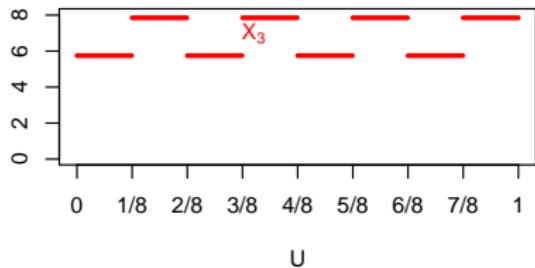


X_2



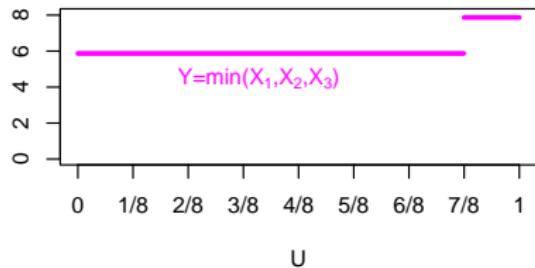
U

X_3



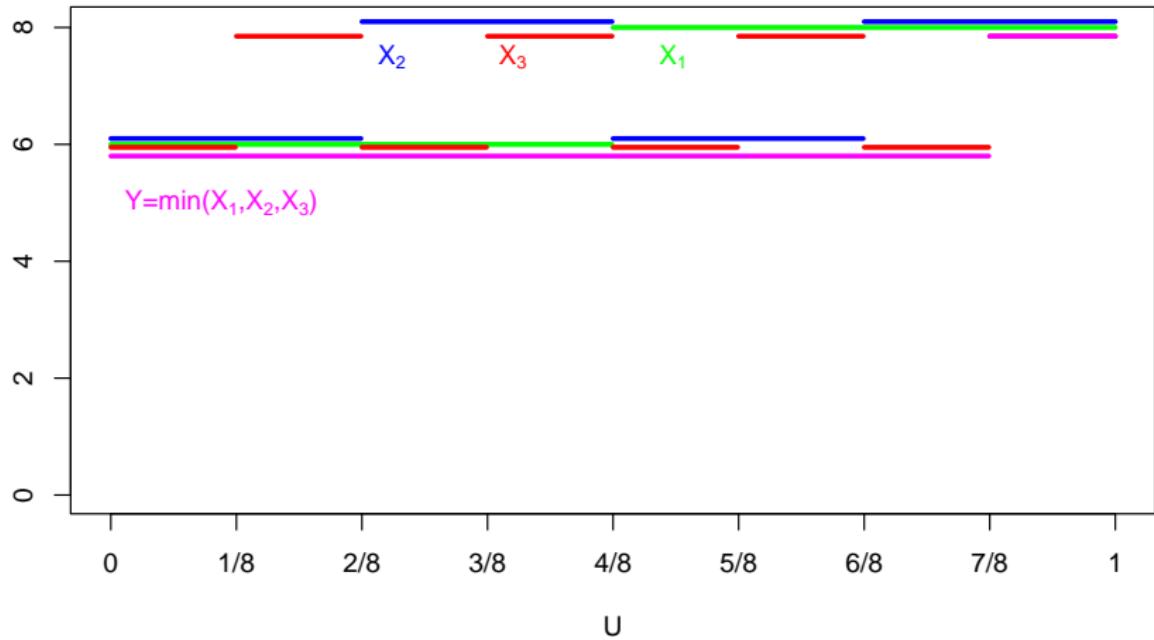
U

$Y = \min(X_1, X_2, X_3)$



U

Graficamos las variables aleatorias, un sólo gráfico.



6. Esperanza

Probabilidades y Estadística (M)

María Eugenia Szretter Noste

Departamento de Matemática e
Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Primer cuatrimestre 2020



Ejemplo 6.1 (leonas)

Una leona da a luz entre 1 y 4 cachorros por camada. Sea X la cantidad de crías que tiene una leona en una camada, la función de probabilidad puntual de X está dada por

k	1	2	3	4
$p_X(k)$	0.2	0.3	0.1	0.4

Un grupo de estudio de ecología animal selecciona al azar n partos de leona, y registra lo obtenido. Sean X_1, X_2, \dots, X_n los tamaños de las camadas observadas. Este grupo está interesado en averiguar la cantidad promedio de leones que nacen por camada, es decir en $\frac{1}{n} \sum_{i=1}^n X_i$. Para fijar ideas, toma $n = 30$. Los valores obtenidos en una realización del experimento resultan ser:

1 4 3 2 4 4 4 4 4 3 1 4 4 1 1 4 1 4 3 1 2 4 1 4 2 2 4 2 1 1

Ejemplo 6.1, cont.

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n X_i &= \frac{1}{30} \{1 + 4 + 3 + \cdots 2 + 1 + 1\} \\&= \frac{1}{30} \left\{ \underbrace{1 + \cdots + 1}_{n_1} + \underbrace{2 + \cdots + 2}_{n_2} + \underbrace{3 + \cdots + 3}_{n_3} + \underbrace{4 + \cdots + 4}_{n_4} \right\} \\&= \frac{1}{30} \{1 \cdot n_1 + 2 \cdot n_2 + 3 \cdot n_3 + 4 \cdot n_4\} \\&= 1 \cdot \frac{n_1}{n} + 2 \cdot \frac{n_2}{n} + 3 \cdot \frac{n_3}{n} + 4 \cdot \frac{n_4}{n} \underset{n \rightarrow \infty}{\approx} \sum_{k=1}^4 k p_X(k) \\&= \sum_{k \in R_X} k p_X(k)\end{aligned}$$

6. Motivación

Consideremos una variable aleatoria discreta X tomando los valores $\{k : k \in R_X\}$ y con función de probabilidad puntual $p_X(k)$, de forma tal que $0 \leq p_X(k) \leq 1$ y $\sum_{k \in R_X} p_X(k) = 1$. Sean X_1, X_2, \dots, X_n n repeticiones independientes de nuestra variable aleatoria. Si denotamos por n_k a la cantidad de veces que observamos el resultado k en las n repeticiones, tenemos que

$$\frac{1}{n} \sum_{i=1}^n X_i = \sum_{k \in R_X} k \frac{n_k}{n}.$$

Cuando el número de repeticiones n tiende a ∞ , sabemos que n_k/n converge a $p_X(k)$. Tenemos entonces que el *promedio muestral de n repeticiones* de nuestra variable aleatoria converge a $\sum_{k \in R_X} k p_X(k)$, como vimos en el ejemplo.

Definición 6.1 (esperanza de X discreta)

Dada una variable aleatoria discreta X tal que $\sum_{k \in R_X} |k| p_X(k) < \infty$. Definimos la **esperanza, o valor medio de X** mediante la fórmula,

$$E[X] = \sum_{k \in R_X} k p_X(k). \quad (1)$$

Requerimos que el valor absoluto de la serie converja para que el valor de la $E(X)$ no cambie por reordenamientos de los k . Cuando al menos una de estas sumas

$$\sum_{k \in R_X, k > 0} k p_X(k) < \infty \quad \text{ó} \quad \sum_{k \in R_X, k < 0} k p_X(k) > -\infty.$$

es finita, la suma (1) está bien definida, aunque puede tomar el valor $\pm\infty$.

Ejemplo 6.1, cont.

Sea X la cantidad de crías que tiene una leona en una camada, la función de probabilidad puntual de X está dada por

k	1	2	3	4
$p_X(k)$	0.2	0.3	0.1	0.4

Entonces,

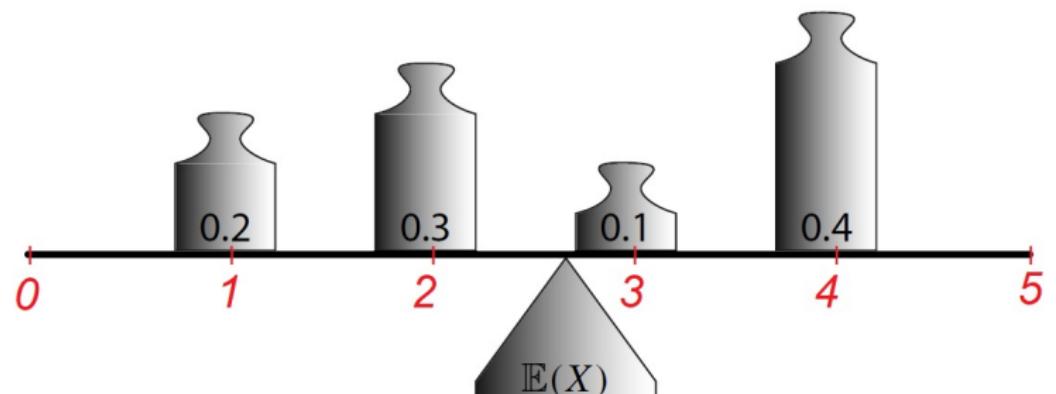
$$\begin{aligned} E[X] &= \sum_{k \in R_X} k p_X(k) = 1 \cdot 0,2 + 2 \cdot 0,3 + 3 \cdot 0,1 + 4 \cdot 0,4 \\ &= 0,2 + 0,6 + 0,3 + 1,6 = 2,7 \end{aligned}$$

Observemos que en este caso, $E(X)$ ¡no está en el rango de X ! La esperanza no necesita ser uno de los valores posibles para la variable aleatoria.

Esperanza como centro de masa

El concepto de esperanza es análogo al concepto físico del centro de gravedad de una distribución de masa (o centro de masa). Para la variable X que toma los valores k_i con probabilidad $p_X(k_i)$, imaginamos el eje x como una barra que no tiene masa y sobre él ubicamos en los puntos k_i un peso con masa $p_X(k_i)$ (masa total 1), entonces el punto en el cual al apoyar la barra ésta estaría en equilibrio se conoce como el centro de gravedad.

Ejemplos 6.1, cont.



Ejemplo

Ejemplo 6.2 (distribución Bernoulli)

Si $X \sim Be(p)$, entonces $E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$

Ejemplo 6.3 (función indicadora)

Es un caso particular del ejemplo anterior, si $X = I_A$ con $A \in \mathcal{F}$, entonces $E(X) = P(A)$.

Ejemplo 6.4 (distribución geométrica)

Sea $X \sim \mathcal{G}(p)$, es decir, $R_X = \mathbb{N}$ y $p_X(k) = p(1 - p)^{k-1}$, para $k \in \mathbb{N}$.

Hallemos su esperanza.

$$E[X] = \sum_{k \in R_X} k p_X(k) = \sum_{k=1}^{+\infty} kp(1 - p)^{k-1}$$

Como $kq^{k-1} = \frac{d}{dq} q^k$, y dentro del radio de convergencia de una serie de potencias podemos intercambiar las operaciones de suma y diferenciación, obtenemos

$$\sum_{k=1}^{+\infty} kq^{k-1} = \sum_{k=1}^{\infty} \frac{d}{dq} [q^k] = \frac{d}{dq} \left[\sum_{k=1}^{\infty} q^k \right] = \frac{d}{dq} \left[\frac{q}{1-q} \right] = \frac{1}{(1-q)^2}, \text{ resulta}$$

$$E(X) = \sum_{k=1}^{+\infty} kp(1 - p)^{k-1} = \frac{1}{p}.$$

Por ejemplo, si el 10 % de los items son defectuosos, debemos revisar en promedio 10 items para encontrar uno defectuoso, como podríamos haber anticipado.

Ejemplo 6.5 (v.a. con esperanza infinita)

Sea X variable aleatoria con $R_X = \mathbb{N}$ y $p_X(k) = c \frac{1}{k^2}$ con c para que las puntuales sumen uno. Entonces

$\sum_{k \in R_X} k p_X(k) = \sum_{k=1}^{+\infty} k c \frac{1}{k^2} = c \sum_{k=1}^{+\infty} \frac{1}{k} = +\infty$. Podemos aceptar valores $+\infty$ y decir que en este caso, $E[X] = +\infty$.

Ejemplo 6.6 (v.a. sin esperanza)

Sea X variable aleatoria con $R_X = \mathbb{Z} - \{0\}$ y $p_X(k) = a \frac{1}{k^2}$ con a para que las puntuales sumen uno. Entonces

$$\sum_{k \in R_X: k > 0} k p_X(k) = \sum_{k=1}^{+\infty} k a \frac{1}{k^2} = a \sum_{k=1}^{+\infty} \frac{1}{k} = +\infty \text{ y}$$

$\sum_{k \in R_X: k < 0} k p_X(k) = \sum_{k=1}^{+\infty} (-k) a \frac{1}{k^2} = -a \sum_{k=1}^{+\infty} \frac{1}{k} = -\infty$. Entonces no podemos definir bien $\sum_{k \in R_X} k p_X(k)$ porque la suma no converge absolutamente, y esta distribución no tiene esperanza.

Caso continuo

Definición 6.2

Dada una variable aleatoria continua $X : (\Omega, \mathcal{F}, P) \rightarrow \mathbb{R}$ con función de densidad $f_X(x)$, definimos la esperanza de X mediante la fórmula,

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx , \quad (2)$$

siempre que $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$.

La condición $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$ es equivalente a la condición discreta $\sum_{k \in R_X} |k| p_X(k) < \infty$. Si $f_X(x) = 0$ para $x \leq 0$ y $\int_{-\infty}^{\infty} x f_X(x) dx = +\infty$ podemos decir que $E[X] = +\infty$, con una conclusión similar para el caso en que $f_X(x) = 0$ para $x \geq 0$, con $E[X] = -\infty$.

Ejemplo 6.7

Si $X \sim \mathcal{U}(0, 1)$, $E(X) = \int_0^1 x dx = \frac{1}{2}$

Para entender la definición (2), veamos a las variables (absolutamente) continuas como límites de discretas. Sea X absolutamente continua con $f_X(x) = 0$ para $x \notin [a, b]$. Dado $n \in \mathbb{N}$, dividimos al intervalo $[a, b]$ en n

subintervalitos, con extremos en los puntos

$$a + \frac{(b-a)}{n} j$$

, con

$j = 0, \dots, n$, y sea $X_n : (\Omega, \mathcal{F}, P) \rightarrow \mathbb{R}$ la variable aleatoria discreta dada por

$$X_n(\omega) = \sum_{j=0}^{n-1} \left[a + \frac{(b-a)}{n} j \right] I_{[a + \frac{(b-a)}{n} j, a + \frac{(b-a)}{n} (j+1)]}(X(\omega))$$

O sea,

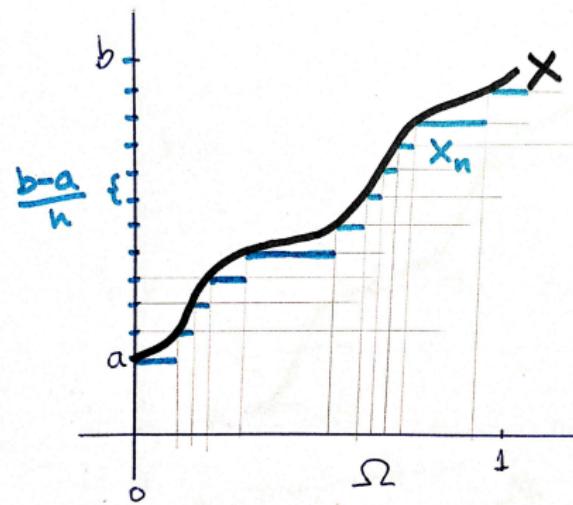
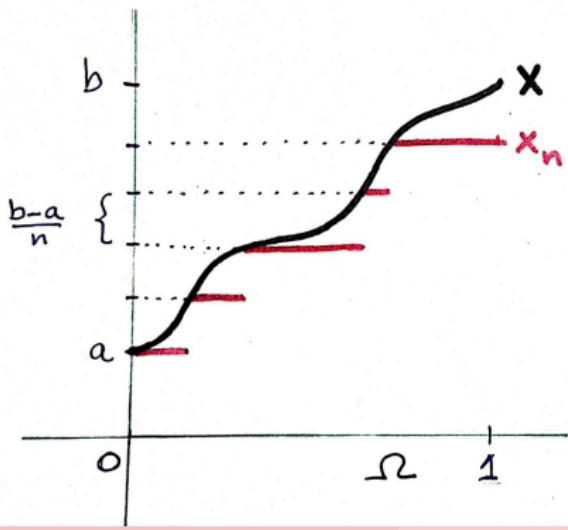
$$X_n(\omega) = a + \frac{(b-a)}{n} j \quad \text{si } a + \frac{(b-a)}{n} j \leq X(\omega) \leq a + \frac{(b-a)}{n} (j+1)$$

para $j = 0, \dots, (n-1)$

$$X_n(\omega) = a + \frac{(b-a)}{n} j$$

$$\text{si } a + \frac{(b-a)}{n} j \leq X(\omega) \leq a + \frac{(b-a)}{n} (j+1)$$

Figura 1: Representación esquemática de la variable aleatoria X con $\Omega = [0, 1]$. La X continua en negro, la discreta X_n que la aproxima en rosa a la izq para $n = 5$, en azul para $n = 11$ a la derecha



$$X_n(\omega) = a + \frac{(b-a)}{n} j$$

$$\text{si } a + \frac{(b-a)}{n} j \leq X(\omega) \leq a + \frac{(b-a)}{n} (j+1)$$

Entonces,

$$E[X_n] = \sum_{k \in R_{X_n}} k p_{X_n}(k)$$

$$= \sum_{j=0}^{n-1} \left(a + \frac{(b-a)}{n} j \right) P \left(X \in \left[a + \frac{(b-a)}{n} j, a + \frac{(b-a)}{n} (j+1) \right] \right)$$

$$= \sum_{j=0}^{n-1} \left(a + \frac{(b-a)}{n} j \right) \int_{a + \frac{(b-a)}{n} j}^{a + \frac{(b-a)}{n} (j+1)} f_X(x) dx$$

$$\approx \sum_{j=0}^{n-1} \int_{a + \frac{(b-a)}{n} j}^{a + \frac{(b-a)}{n} (j+1)} x f_X(x) dx = \int_a^b x f_X(x) dx$$

Funciones de variables o vectores aleatorios

En general $E[g(X)] \neq g(E[X])$

Ejemplo 6.8

Sea X que toma los valores $-1, 0, 1$ con probabilidades $\frac{1}{3}$ cada uno, tenemos

$$E[X] = 0$$

Sea $Y = X^2$, toma valores 0 y 1 , con probabilidades $\frac{1}{3}$ y $\frac{2}{3}$ respectivamente, tenemos

$$E[Y] = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3} \neq 0^2$$

Teorema 6.1

Sea (X_1, X_2, \dots, X_n) un vector aleatorio n -dimensional. Sea $g: \mathbb{R}^n \rightarrow \mathbb{R}$ una función boreiana. Consideremos la variable aleatoria $Y = g(X_1, X_2, \dots, X_n)$. Entonces:

- ① Si (X_1, X_2, \dots, X_n) es discreto y

$\sum_{x_1, x_2, \dots, x_n} |g(x_1, x_2, \dots, x_n)| p_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) < +\infty$
entonces

$$E[g(X_1, \dots, X_n)] = \sum_{x_1, \dots, x_n} g(x_1, \dots, x_n) p_{X_1 X_2 \dots X_n}(x_1, \dots, x_n)$$

- ② Si (X_1, X_2, \dots, X_n) es continuo y

$\int \dots \int |g(x_1, \dots, x_n)| f_{X_1 X_2 \dots X_n}(x_1, \dots, x_n) dx_1 dx_2 \dots dx_n < +\infty$
entonces

$$E[g(X_1, \dots, X_n)] = \int \dots \int g(x_1, \dots, x_n) f_{X_1 \dots X_n}(x_1, \dots, x_n) dx_1 \dots dx_n$$

Demostración. (caso discreto).

Llamamos $\tilde{\mathbf{X}} = (X_1, \dots, X_n)$, $Y = g(X_1, X_2, \dots, X_n) = g(\tilde{\mathbf{X}})$

$$E(g(\tilde{\mathbf{X}})) = E(Y) = \sum_{y \in R_Y} y p_Y(y)$$

$$\begin{aligned} &= \sum_{y \in R_Y} y P(g(\tilde{\mathbf{X}}) = y) = \sum_{y \in R_Y} y \left[\sum_{\mathbf{x} \in R_{\tilde{\mathbf{X}}} : g(\mathbf{x}) = y} P(\tilde{\mathbf{X}} = \mathbf{x}) \right] \\ &= \sum_{y \in R_Y} \sum_{\mathbf{x} \in R_{\tilde{\mathbf{X}}} : g(\mathbf{x}) = y} y P(\tilde{\mathbf{X}} = \mathbf{x}) = \sum_{y \in R_Y} \sum_{\mathbf{x} \in R_{\tilde{\mathbf{X}}} : g(\mathbf{x}) = y} g(\mathbf{x}) P(\tilde{\mathbf{X}} = \mathbf{x}) \\ &= \sum_{\mathbf{x} \in R_{\tilde{\mathbf{X}}}} g(\mathbf{x}) P(\tilde{\mathbf{X}} = \mathbf{x}) \end{aligned}$$

El argumento para probar el caso continuo es análogo, pero para que la prueba sea rigurosa hay que usar argumentos de integración que se ven en teoría de la medida.

Aunque forma parte del resultado anterior, para enfatizarlo, escribamos el enunciado en el caso de $n = 1$.

Teorema 6.2

Sea X una variable aleatoria. Sea $g : \mathbb{R} \rightarrow \mathbb{R}$ boreiana y definamos $Y = g(X)$.

- ① Si X es discreta y $\sum_{k \in R_X} |g(k)| p_X(k) < +\infty$, entonces

$$E[Y] = \sum_{k \in R_X} g(k) p_X(k)$$

- ② Si X es continua y $\int_{-\infty}^{+\infty} |g(x)| f_X(x) dx < +\infty$, entonces

$$E[Y] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx.$$

Ejemplo 6.8, continuación

Sea X que toma los valores $-1, 0, 1$ con probabilidades $\frac{1}{3}$ cada uno, tenemos

$$E[X] = 0$$

Sea $Y = X^2 = g(X)$, toma valores 0 y 1 , con probabilidades $\frac{1}{3}$ y $\frac{2}{3}$ respectivamente, Habíamos calculado la $E[Y]$ calculando primero la p_Y . Hagámoslo ahora directamente usando la fórmula del Teorema tenemos

$$E[Y] = g(-1) \cdot \frac{1}{3} + g(0) \cdot \frac{1}{3} + g(1) \cdot \frac{1}{3} = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

Ejemplo 3.2, dardos

Se elige un punto al azar sobre un tablero circular de radio siete, y definamos la variable aleatoria D que mide la distancia al centro del tablero. Nos interesa calcular $E(D)$, el valor esperado de la distancia al centro del tablero.

Sea el vector aleatorio (X, Y) con distribución $\mathcal{U}(A)$, la distribución uniforme en $A = \{(x, y) : x^2 + y^2 \leq 49\}$ la bola de radio 7 en \mathbb{R}^2 . Luego $f_{XY}(x, y) = \frac{1}{49\pi} I_A(x, y)$ Entonces, $D = g(X, Y) = \sqrt{X^2 + Y^2}$, y por el Teorema 6.1 tenemos

$$\begin{aligned} E[D] &= E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy \\ &= \int \int_A \sqrt{x^2 + y^2} \frac{1}{49\pi} dx dy \end{aligned}$$

Haciendo el cambio a polares, $x = r \cos(\theta)$, $y = r \sin(\theta)$, con jacobiano igual a r . La región A está dada por, $A = \{(r, \theta), 0 \leq r \leq 7, 0 \leq \theta < 2\pi\}$.

Ejemplo 3.2, dardos

$$\begin{aligned} E[D] &= \int \int_A \sqrt{x^2 + y^2} \frac{1}{49\pi} dx dy \\ &= \int_0^{2\pi} \int_0^7 \frac{1}{49\pi} r^2 dr d\theta = \frac{1}{49\pi} 2\pi \frac{r^3}{3} \Big|_0^7 = \frac{14}{3} = 4,66667 \end{aligned}$$

Otra forma de resolver este ejercicio hubiera sido hallar la densidad de D , completando el vector aleatorio (D, V) con otra variable V de modo que se cumplan las hipótesis del teorema de cambio de variables, calcular la densidad f_{DV} , luego la marginal f_D y finalmente con ella la esperanza de D .

Ejemplo 6.9

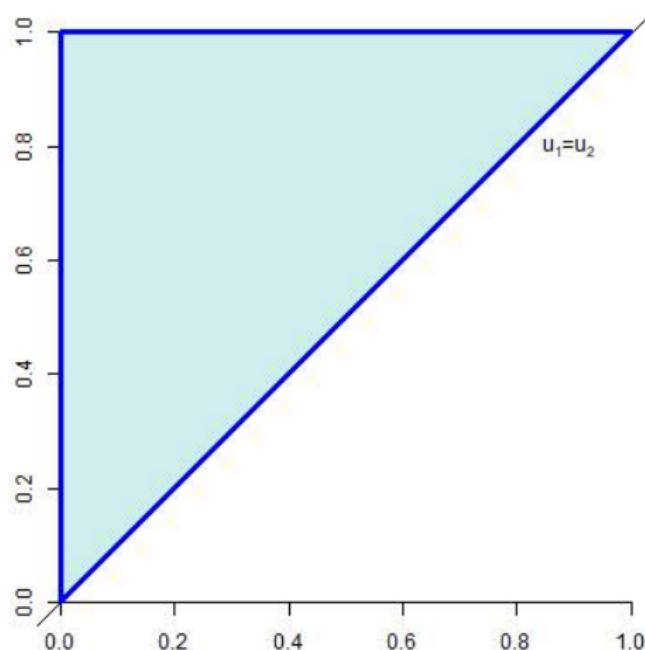
Un segmento de longitud uno se corta en dos puntos elegidos al azar. Hallar la longitud esperada del segmento central.

Interpretamos que esta pregunta significa que las ubicaciones de los dos puntos de ruptura son variables aleatorias uniformes independientes U_1 y U_2 . Por lo tanto, necesitamos calcular $E|U_1 - U_2|$. El Teorema 6.1 nos dice que no necesitamos encontrar la función de densidad de $|U_1 - U_2|$ sino que simplemente debemos integrar la función $|u_1 - u_2|$ multiplicada por la densidad conjunta de (U_1, U_2) , (que será el producto de las dos densidades marginales puesto que U_1 y U_2 son independientes) , $f_{U_1 U_2}(u_1, u_2) = 1 \cdot I_{[0,1]}(u_1)I_{[0,1]}(u_2)$. Entonces

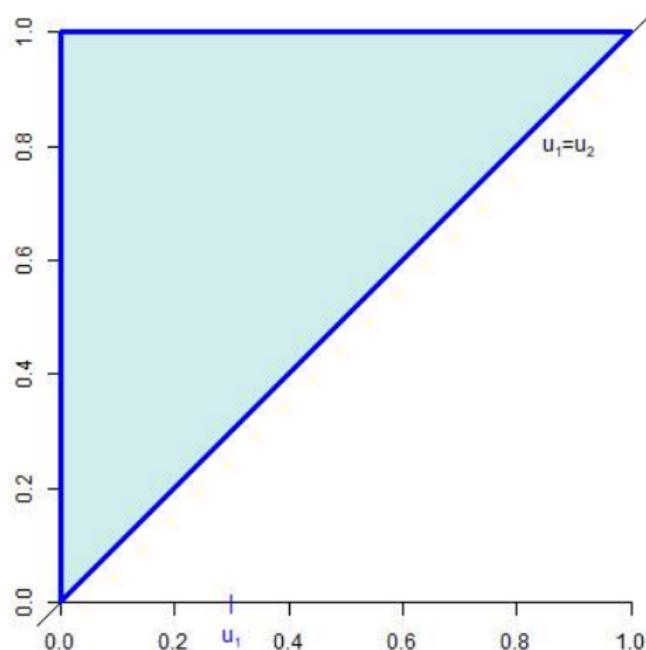
$$E|U_1 - U_2| = \int_0^1 \int_0^1 |u_1 - u_2| du_1 du_2$$

Dividimos esta integral en dos: sobre el triángulo inferior y superior del $[0, 1] \times [0, 1]$ dividido por la diagonal $u_1 = u_2$.

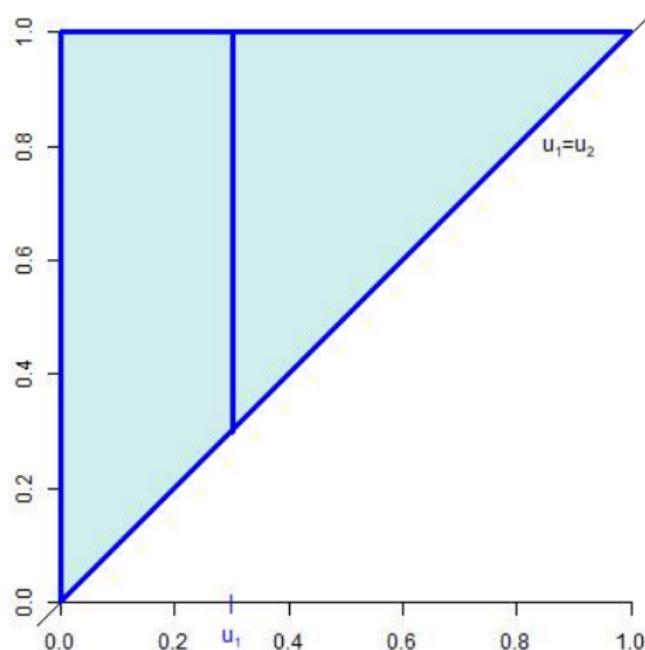
Ejemplo 6.9, cont.



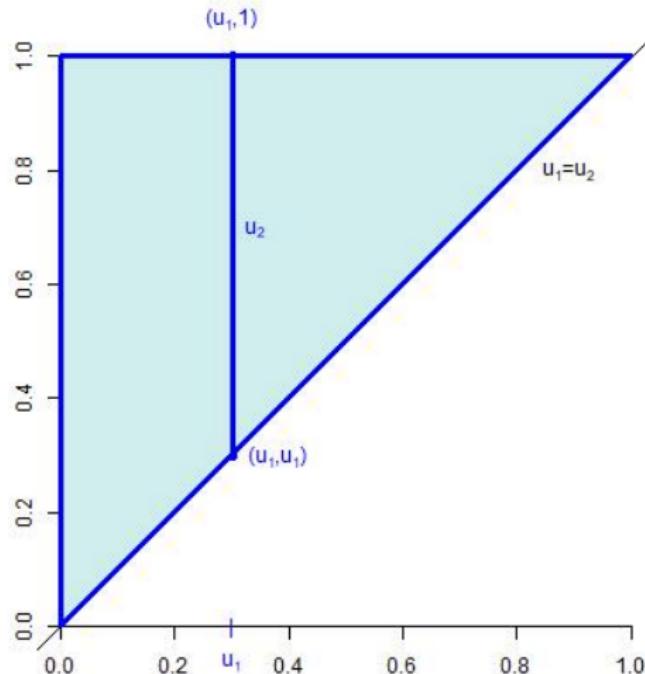
Ejemplo 6.9, cont.



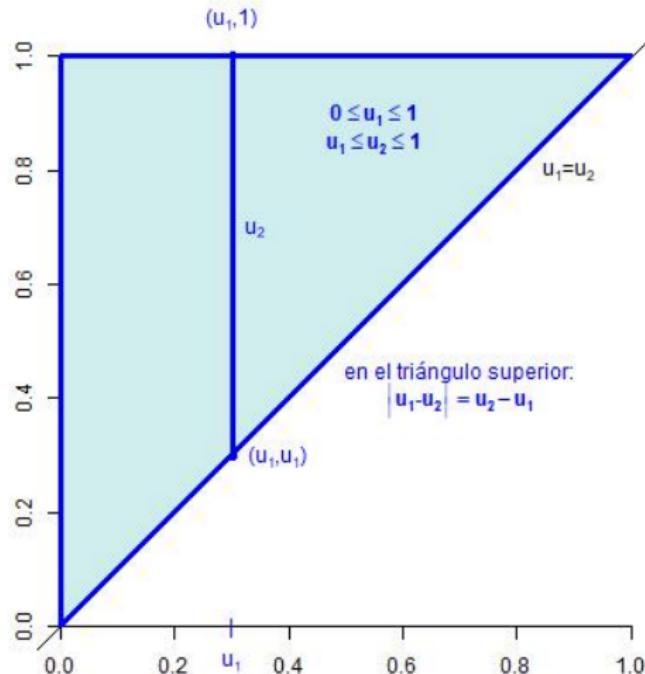
Ejemplo 6.9, cont.



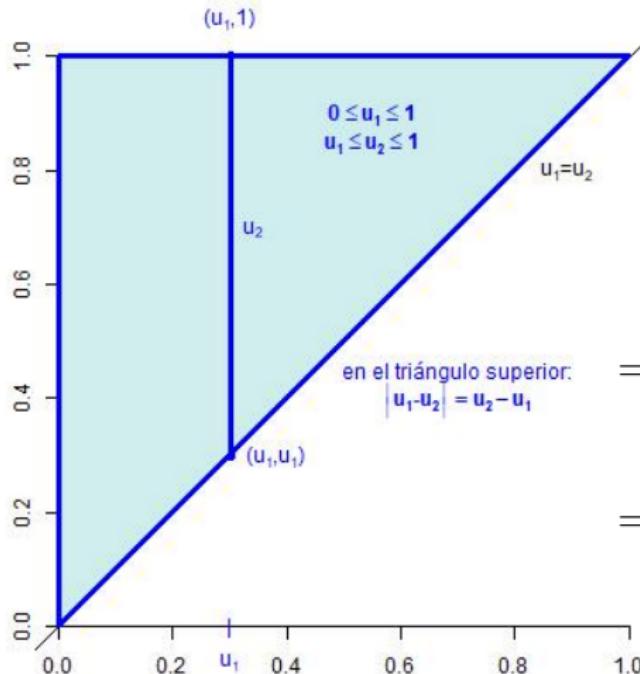
Ejemplo 6.9, cont.



Ejemplo 6.9, cont.

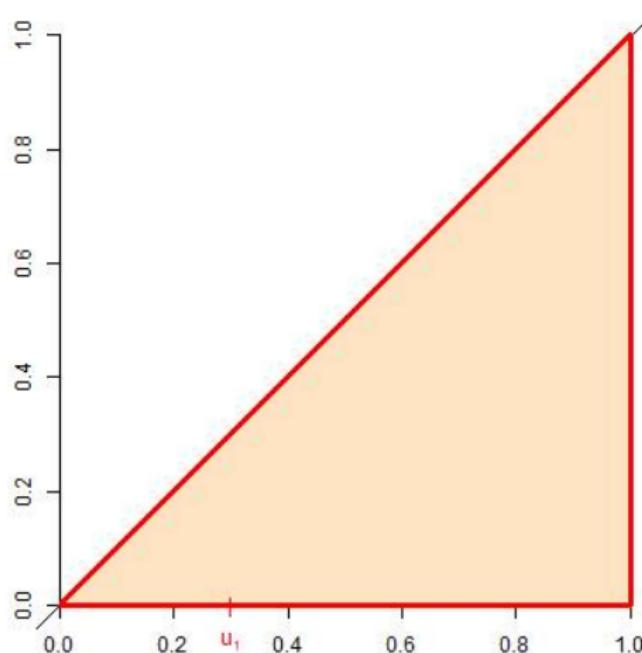


Ejemplo 6.9, cont.

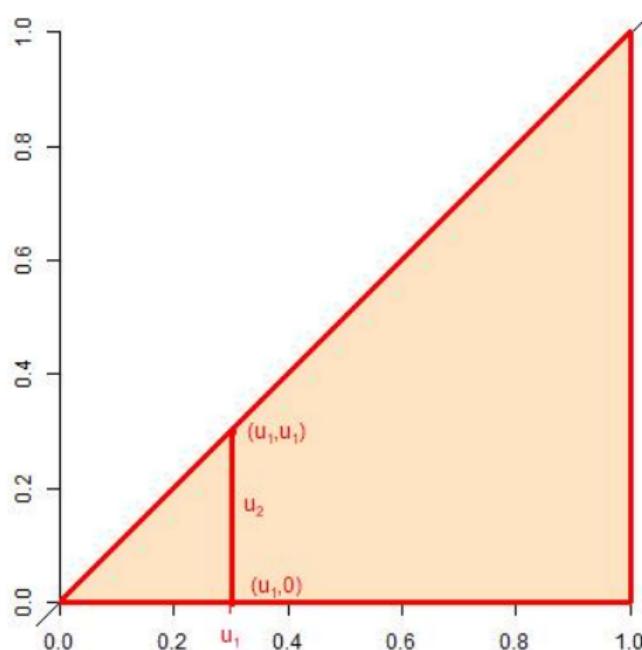


$$\int \int_{\text{triang sup}} |u_1 - u_2| du_1 du_2$$
$$= \int \int_{\text{triang sup}} (u_2 - u_1) du_1 du_2$$
$$= \int_0^1 \int_{u_1}^1 (u_2 - u_1) du_2 du_1$$

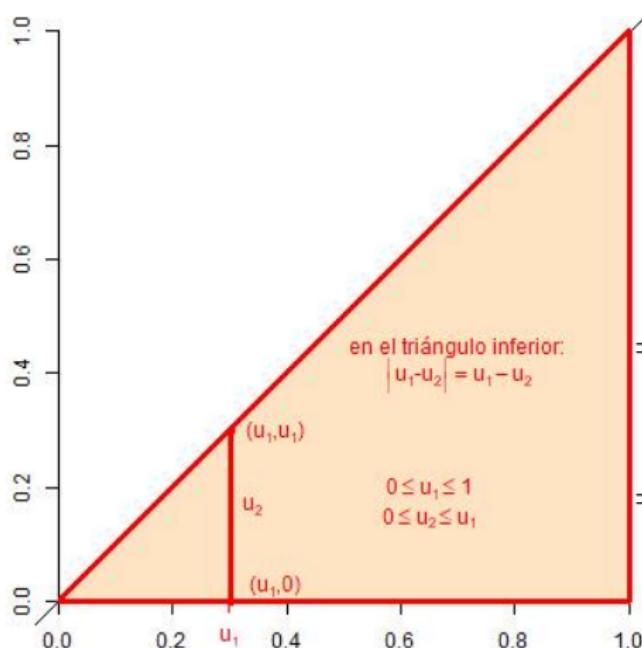
Ejemplo 6.9, cont.



Ejemplo 6.9, cont.



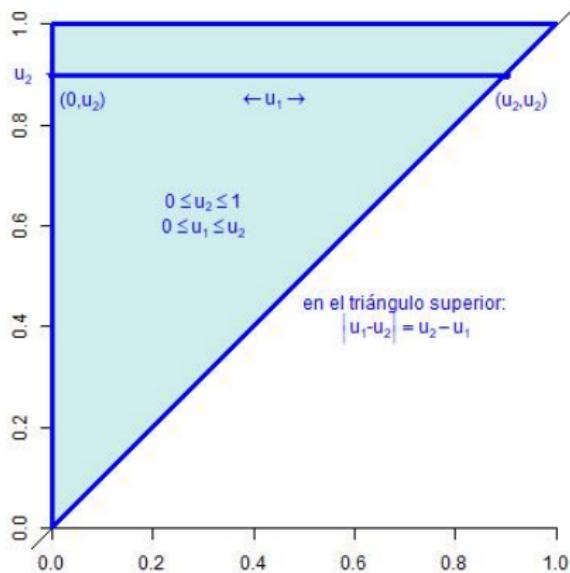
Ejemplo 6.9, cont.



$$\begin{aligned} & \int \int |u_1 - u_2| du_1 du_2 \\ & \text{triang inf} \\ &= \int \int (u_1 - u_2) du_1 du_2 \\ & \text{triang inf} \\ &= \int_0^1 \int_0^{u_1} (u_1 - u_2) du_2 du_1 \end{aligned}$$

Ejemplo 6.9, cont.

$$E |U_1 - U_2| = \int_0^1 \int_0^{u_1} (u_1 - u_2) du_2 du_1 + \int_0^1 \int_{u_1}^1 (u_2 - u_1) du_2 du_1$$



La segunda integral es sobre el triángulo superior, al cual podemos escribir como $\{(u_1, u_2) : 0 \leq u_2 \leq 1, 0 \leq u_1 \leq u_2\}$, luego nos da $\int_0^1 \int_0^{u_2} (u_2 - u_1) du_1 du_2$ que es igual a la otra integral (la roja) pero con los subíndices cambiados de lugar.

Ejemplo 6.9, cont.

Finalmente

$$\begin{aligned} E|U_1 - U_2| &= 2 \int_0^1 \int_0^{u_1} (u_1 - u_2) du_2 du_1 = 2 \int_0^1 \left(u_1 u_2 - \frac{u_2^2}{2} \right) \Big|_{u_2=0}^{u_2=u_1} du_1 \\ &= \int_0^1 u_1^2 du_1 = \frac{1}{3} \end{aligned}$$

Lo cual puede interpretarse desde el punto de vista intuitivo: cada intervalo debiera tener la misma longitud en promedio, o sea $\frac{1}{3}$ cada uno.

Propiedades de la esperanza

Lema 6.3 (linealidad de la esperanza)

Sean X e Y variables aleatorias definidas en (Ω, \mathcal{F}, P) con esperanza finita. Entonces

- a) $E[aX] = aE[X]$, para todo $a \in \mathbb{R}$.
- b) $E|X + Y| \leq E|X| + E|Y|$.
- c) $E[X + Y] = E[X] + E[Y]$.

En general para X_1, \dots, X_n variables aleatorias con esperanza finita y constantes $a_1, \dots, a_n \in \mathbb{R}$ tenemos $E[\sum_{i=1}^n a_i X_i] = \sum_{i=1}^n a_i E[X_i]$.

Demostración. (caso discreto).

- a) Sea $Z = aX$, $p_Z(z) = p_X(x)$ si $z = ax$, $R_Z = aR_X$, resulta

$$E[aX] = \sum_{z \in R_Z} z p_Z(z) = \sum_{x \in R_X} a x p_X(x) = a \sum_{x \in R_X} x p_X(x) = a E[X].$$

Demostración. (caso discreto).

Para el b), sea $Z = X + Y$.

$$\begin{aligned} E|Z| &= \sum_{z \in R_Z} |z| p_Z(z) = \sum_{z \in R_Z} |z| \sum_{\substack{x \in R_X, y \in R_Y \\ x+y=z}} p_{XY}(x, y) \\ &= \sum_{z \in R_Z} \sum_{\substack{x \in R_X, y \in R_Y \\ x+y=z}} |x+y| p_{XY}(x, y) = \sum_{x \in R_X} \sum_{y \in R_Y} |x+y| p_{XY}(x, y) \\ &\leq \sum_{x \in R_X} \sum_{y \in R_Y} (|x| + |y|) p_{XY}(x, y) \\ &= \sum_{x \in R_X} \sum_{y \in R_Y} |x| p_{XY}(x, y) + \sum_{x \in R_X} \sum_{y \in R_Y} |y| p_{XY}(x, y) \\ &= \sum_{x \in R_X} |x| \underbrace{\sum_{y \in R_Y} p_{XY}(x, y)}_{p_X(x)} + \sum_{y \in R_Y} |y| \underbrace{\sum_{x \in R_X} p_{XY}(x, y)}_{p_Y(y)} = E|X| + E|Y| \end{aligned}$$



Demostración.

Los cambios de orden de suma están justificados porque las series convergen absolutamente. Luego, el ítem b) prueba además que la esperanza de $X + Y$ es finita, por serlo la del $|X + Y|$.

La prueba de c) es la misma cuenta que el ítem b), sin usar las barras de módulo y con una igualdad en el lugar en que la prueba del ítem b) lleva un \leq .

Los items a) y c) dan la linealidad de la esperanza. □

El Lema 6.3, es decir, la linealidad de la esperanza, es una propiedad muy útil. Permite, por ejemplo, calcular la esperanza de la distribución binomial y de la binomial negativa de forma sencilla.

Propiedades de la esperanza

Lema 6.4 (monotonía de la esperanza)

Sean X e Y variables aleatorias definidas en (Ω, \mathcal{F}, P) con esperanza finita. Entonces si $X \leq Y$, es decir, $X(\omega) \leq Y(\omega) \forall \omega \in \Omega$, luego $E[X] \leq E[Y]$.

Demostración. (caso discreto).

$X \leq Y \Leftrightarrow 0 \leq Y - X = Z$. Como $R_Z \subset \mathbb{R}_{\geq 0}$, resulta $E[Z] \geq 0$.

$$E[Z] = E[Y - X] \stackrel{\text{linealidad}}{=} E[Y] - E[X] \geq 0, \Leftrightarrow E[Y] \geq E[X]. \quad \square$$

Lema 6.5 (esperanza de una constante)

Sean $X = a$ con probabilidad uno, es decir, sea X una variable aleatoria constante, entonces $E[X] = a$.

Demostración.

$$p_X(a) = 1, \text{ luego } E[X] = a \cdot p_X(a) = a \quad \square$$

Propiedades de la esperanza

Lema 6.6 (módulo)

$$|E(X)| \leq E(|X|)$$

Demostración.

(caso continuo)

$$|E(X)| = \left| \int xf_X(x)dx \right| \leq \int |x| f_X(x)dx = E(|X|),$$

donde la última igualdad es válida por el Teorema 6.1.



Propiedades de la esperanza

Lema 6.7

Si $X \geq 0$ y $E[X] = 0$ entonces $P(X = 0) = 1$.

Demostración.

Para cada $n \in \mathbb{N}$, definimos la v.a. $Y_n(\omega) = \frac{1}{n} I_{[\frac{1}{n}, +\infty)}(X(\omega))$, entonces $Y_n \leq X$. Luego, como Y_n es discreta, tenemos la igualdad

$$E(Y_n) = \frac{1}{n} P\left(X \geq \frac{1}{n}\right) \underbrace{\leq}_{\text{monotonía de } E} E(X) = 0.$$

Luego, $P\left(X \geq \frac{1}{n}\right) = 0$ para todo $n \in \mathbb{N}$ y como la sucesión creciente de eventos $\{X \geq \frac{1}{n}\}$ converge a $\{X > 0\}$, es decir,
 $\bigcup_{n \geq 1} \{X \geq \frac{1}{n}\} = \{X > 0\}$, por la monotonía de la probabilidad tenemos que $\lim_{n \rightarrow +\infty} P\left(X \geq \frac{1}{n}\right) = P(X > 0) = 0$. □

Propiedades de la esperanza

En general no es cierto que si X e Y son variables aleatorias con esperanza finita ello implique que la $E[XY]$ también lo sea. Esto sí ocurre en el caso en el que sean independientes, donde además se verifica la siguiente propiedad.

Lema 6.8 (esperanza de variables independientes)

Sean X e Y variables aleatorias independientes con esperanza finita.

Entonces la $E[XY]$ también es finita y vale que

- ① $E[XY] = E[X]E[Y]$.
- ② Además $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ para $g, h : \mathbb{R} \rightarrow \mathbb{R}$ borelianás, y tales que $E|g(X)|, E|h(Y)|$ sean finitas.

Propiedades de la esperanza

Demostración.

1) (caso continuo) Usamos el Teorema 6.1 primero para $g(X, Y) = |XY|$, para que probemos que esta esperanza existe.

$$\begin{aligned} E[|XY|] &= E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |xy| f_X(x) f_Y(y) dx dy = \left[\int_{-\infty}^{\infty} |x| f_X(x) dx \right] \left[\int_{-\infty}^{\infty} |y| f_Y(y) dy \right] \\ &= E[|X|]E[|Y|] < +\infty \end{aligned}$$

Ahora, usando el mismo Teorema 6.1 para $g(X, Y) = XY$ y repitiendo los pasos de la cuenta anterior probamos 1). □

Propiedades de la esperanza

Demostración.

- 2) Por el Lema 5.5 (la independencia de variables o vectores aleatorios se mantiene por transformaciones), resultan $g(X)$ y $h(Y)$ independientes, luego, por 1) vale 2). □

Propiedades de la esperanza

Lema 6.9 (relación con la función de distribución)

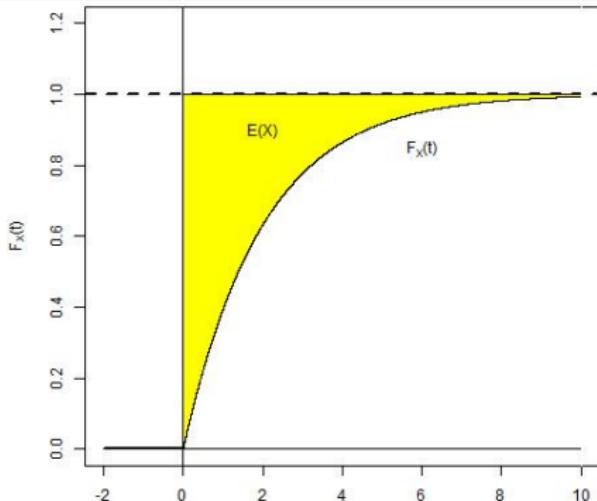
Sea X una variable con esperanza finita, entonces

$$E(X) = \int_0^{+\infty} (1 - F_X(x)) dx - \int_{-\infty}^0 F_X(x) dx \quad (3)$$

Observemos que en el caso en el que $X \geq 0$, la ecuación (3) queda

$$E(X) = \int_0^{+\infty} (1 - F_X(x)) dx$$

y da el área comprendida entre la recta horizontal $y = 1$ y el gráfico de F_X , a la derecha del eje y .



$$E(x) = \int_{-\infty}^{+\infty} x f_x(x) dx = \int_0^{+\infty} \cancel{x} f_x(x) dx + \int_{-\infty}^0 \cancel{x} f_x(x) dx$$

$$= \int_0^{+\infty} \left(\int_0^x dy \right) f_x(x) dx + \int_{-\infty}^0 -\left(\int_x^0 dy \right) f_x(x) dx$$

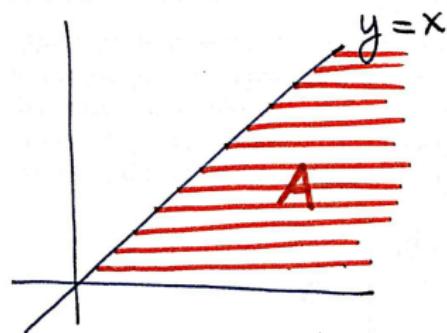
$$= \iint_A f_x(x) dy dx + \iint_B f_x(x) dy dx$$

Cambiamos el orden de integración (Región tipo 1 a tipo 2)

$$\iint_A f_x(x) dx dy = \int_0^{+\infty} \int_y^{+\infty} f_x(x) dx dy$$

A

Cambiamos el orden de integración (tipo I \leftrightarrow tipo II)



$$A = \{(x,y) : 0 < y < x\}$$

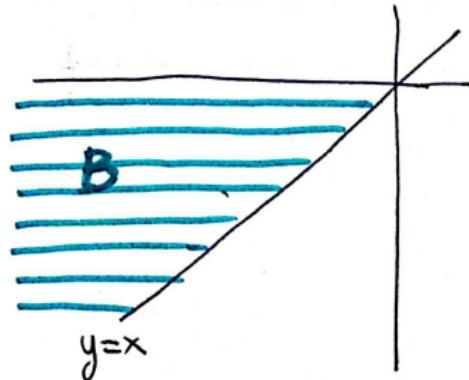
$$= \{(x,y) : y > 0\}$$

$$x > y$$

$$= \int_0^{+\infty} F_x(x) \Big|_{x=y}^{x \rightarrow +\infty} dy = \int_0^{+\infty} [1 - F_x(y)] dy$$

$$\iint_B f_X(x) dx dy = \int_{-\infty}^0 \int_{-\infty}^y f_X(x) dx dy$$

B



$$B = \{(x, y) : 0 > x, x < y < 0\}$$

$$= \left\{ (x, y) : \begin{array}{l} y < 0 \\ x < y \end{array} \right\}$$

$$= \int_{-\infty}^0 F_X(x) \Big|_{x \rightarrow -\infty}^{x=y} dy = \int_{-\infty}^0 F_X(y) dy$$

Luego $E(X) = \int_0^{+\infty} [1 - F_X(y)] dy - \int_{-\infty}^0 F_X(y) dy$

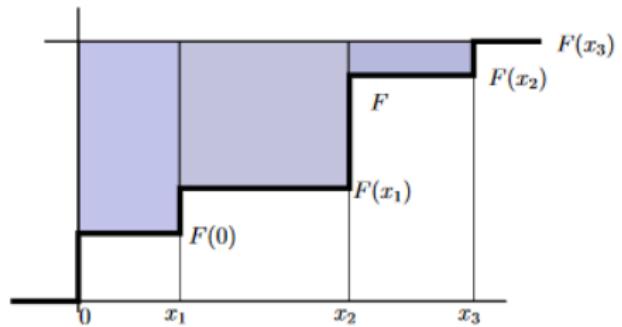
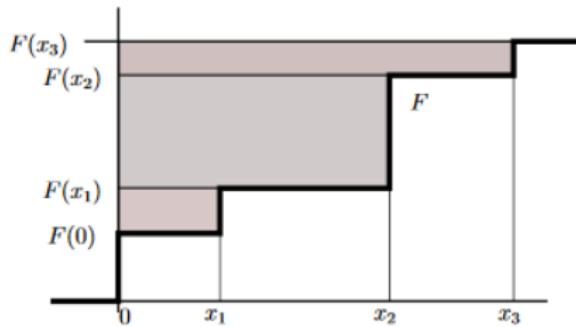
Demostración.

Caso discreto: sólo lo probamos en el caso en que $X \geq 0$ y además podemos escribir su rango $R_X = \{x_0, x_1, x_2, \dots\}$ con $x_0 = 0$ y $x_{i-1} < x_i$ para todo $i \geq 1$. Como $p_X(x_i) = F_X(x_i) - F_X(x_{i-1})$, tenemos

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} x_i p_X(x_i) = \sum_{i=1}^{\infty} \textcolor{red}{x_i} (F_X(x_i) - F_X(x_{i-1})) \\ &= \sum_{i=1}^{\infty} \left[\sum_{j=1}^i (x_j - x_{j-1}) \right] (F_X(x_i) - F_X(x_{i-1})) \quad (\text{cambiamos orden}) \\ &= \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} (x_j - x_{j-1}) (F_X(x_i) - F_X(x_{i-1})) \quad (\text{telescópica}) \\ &= \sum_{j=1}^{\infty} (x_j - x_{j-1}) (1 - F_X(x_{j-1})) = \int_0^{\infty} (1 - F_X(x)) dx \end{aligned}$$

Si 0 no está en el rango, tendremos $F_X(0) = 0$ y exactamente la misma cuenta vale. Cuando X toma valores negativos es análogo al continuo. □

Figura 2: Las figuras (copiadas del apunte de Pablo Ferrari) muestran dos maneras de expresar $\int_0^\infty (1 - F(x))dx$ en el caso discreto. A la izquierda $\sum_{i \geq 1} x_i (F(x_i) - F(x_{i-1}))$ y a la derecha $\sum_{j \geq 1} (x_j - x_{j-1}) (1 - F(x_{j-1}))$



Otra interpretación de la esperanza

Consideremos una variable aleatoria X discreta tomando valores $\{x_1, x_2, \dots, x_m\}$. Suponga que tenemos que poder resumir en un único número a la variable aleatoria. Buscamos entonces la “mejor” constante a que “resuma” o “aproxime” a nuestra variable aleatoria. ¿Qué quiere decir mejor aproxime? ¿Cómo comparamos diferentes valores de a ?

Vamos a considerar una función de pérdida, que mida cuánto *pagamos* al reemplazar la variable aleatoria X por la constante a . Varias son las funciones de pérdida que pueden ser consideradas en probabilidades y estadística, pero una muy habitual es la llamada **pérdida cuadrática**, según la cual el precio que pago al aproximar a X por a cuando $X = x_i$ está dado por $(x_i - a)^2$. Ahora bien, esta pérdida debemos pesarla considerando la frecuencia del valor x_i . Tenemos entonces que **el precio que pagamos al aproximar a la variable aleatoria por la constante a está dado por**

$$H(a) = E[(X - a)^2] = \sum_{i=1}^m (x_i - a)^2 p_X(x_i).$$

$$H(a) = E[(X - a)^2] = \sum_{i=1}^m (x_i - a)^2 p_X(x_i).$$

Buscamos entonces el valor de a que minimiza la función $H(a)$.

Un rápido análisis de la función permite identificar que esta se minimiza cuando $H'(a) = 0$.

$$H'(a) = \sum_{i=1}^m -2(x_i - a) p_X(x_i),$$

por lo tanto, H se minimiza en

$$a = \sum_{i=1}^m x_i p_X(x_i).$$

Tenemos así que $a = E[X]$ es la constante que mejor aproxima a nuestra variable aleatoria X .

Varianza

Acabamos de mostrar que la esperanza es la mejor constante para aproximar a nuestra variable aleatoria. Ahora vamos a querer medir cuál es el precio que pagamos al hacer esta aproximación. Es decir, queremos calcular cuánto vale la función H en $a = E[X]$.

Definición 6.3 (varianza)

Dada una variable aleatoria X con esperanza μ_X , definimos su **varianza** que notaremos $V(X)$ ó $\text{Var}(X)$ o también σ_X^2 mediante la fórmula

$$V(X) = E[(X - \mu_X)^2].$$

La varianza de la variable aleatoria mide cuán dispersa está la variable aleatoria alrededor de su esperanza. Es decir, cuanto más dispersa esté, mayor será el precio que pagaremos al reemplazarla por una constante (peor resumen estaremos haciendo de la variable cuanto menos concentrada esté).

Propiedades de la varianza

Propiedad V.1

$V(X) \geq 0$ y además $V(X) = 0$ si y sólo si la variable aleatoria es constante $X = \mu_X$ con probabilidad 1, o sea, $P(X = \mu_X) = 1$.

Demostración.

Es consecuencia del Lema 6.7



Propiedades de la varianza

La siguiente propiedad, da una fórmula alternativa para computar la varianza de una variable aleatoria.

Propiedad V.2

Sea X una variable aleatoria con $\mu_X = E(X)$, entonces vale que

$$V(X) = E[X^2] - \mu_X^2 .$$

Demostración.

$$\begin{aligned} V(X) &= E[(X - \mu_X)^2] \\ &= E(X^2 - 2\mu_X X + \mu_X^2) \end{aligned}$$

Por la linealidad de la esperanza, resulta

$$\begin{aligned} V(X) &= E(X^2) - 2\mu_X E(X) + \mu_X^2 \\ &= E(X^2) - 2\mu_X^2 + \mu_X^2 \\ &= E(X^2) - \mu_X^2 \end{aligned}$$

Propiedades de la varianza

Propiedad V.3 (cambios de escala)

Dados a y b números reales, tenemos que

$$V(a + bX) = b^2 V(X).$$

Demostración.

Llamemos $Y = a + bX$. Como $E(Y) = a + bE(X)$

$$\begin{aligned} E[(Y - E(Y))^2] &= E\{[a + bX - a - bE(X)]^2\} \\ &= E\{b^2[X - E(X)]^2\} \\ &= b^2 E\{[X - E(X)]^2\} = b^2 V(X) \end{aligned}$$

□

Desvío estándar

Definición 6.4 (desvío estándar)

Dada una variable aleatoria X , definimos el **desvío estándar de X** como la raíz cuadrada de su varianza:

$$DS(X) = \sqrt{V(X)}.$$

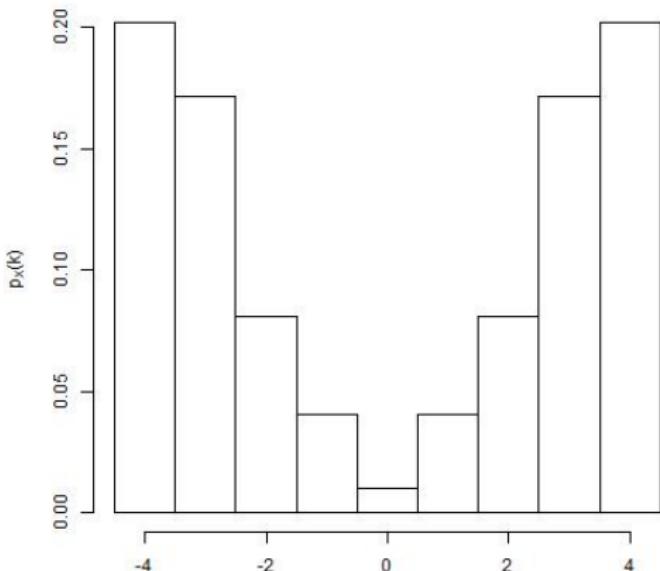
Lo notaremos también σ_X .

Por la **Propiedad V.3 (cambio de escala)** el desvío estándar se transforma de manera natural: si a y b son números reales, e $Y = a + bX$ resulta: $\sigma_Y = |b| \sigma_X$. Luego, si las unidades de medida cambiaron de metros a centímetros, por ejemplo, o sea $Y = 100X$ (X es una medida en metros, Y es la misma medida en cm.) el desvío estándar se multiplicaría por 100, $\sigma_Y = 100\sigma_X$.

Ejemplo: dos variables discretas comparables

X variable discreta con $R_X = \{-4, -3, \dots, 3, 4\}$

k	-4	-3	-2	-1	0	1	2	3	4
$p_X(k)$	0.20	0.17	0.08	0.04	0.01	0.04	0.08	0.17	0.20



Calculamos

$$E[X] = \sum_{k=-4}^4 k p_X(k) = 0$$

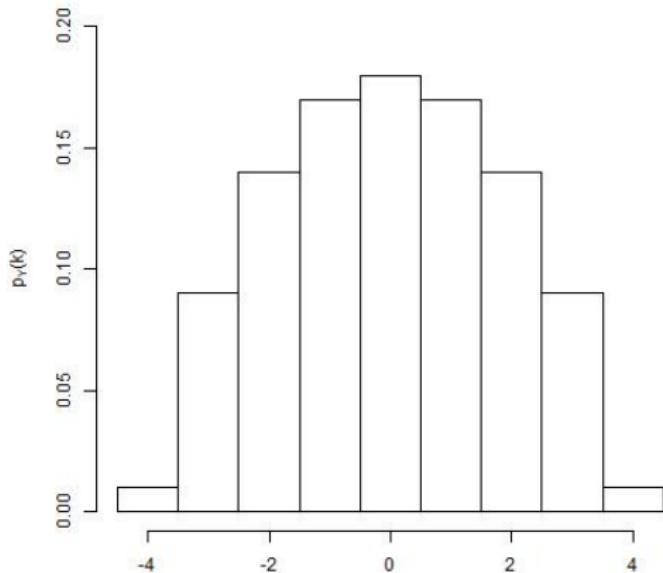
$$E[X^2] = \sum_{k=-4}^4 k^2 p_X(k) = 10,19$$

$$V(X) = E[X^2] - (E[X])^2 = 10,19$$

Ejemplo: dos variables discretas comparables

Y variable discreta con $R_Y = \{-4, -3, \dots, 3, 4\}$

k	-4	-3	-2	-1	0	1	2	3	4
$p_Y(k)$	0.01	0.09	0.14	0.17	0.18	0.17	0.14	0.09	0.01

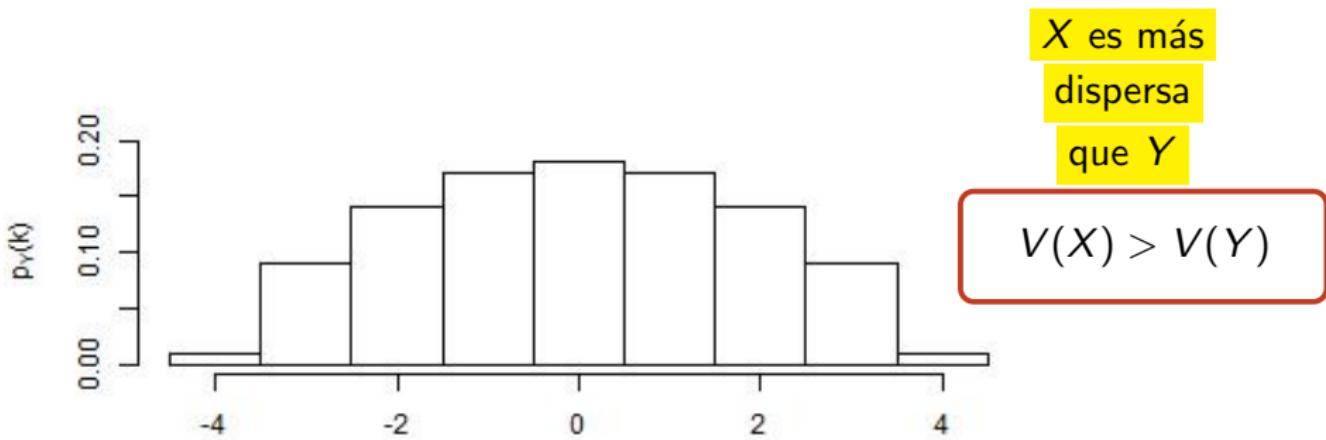
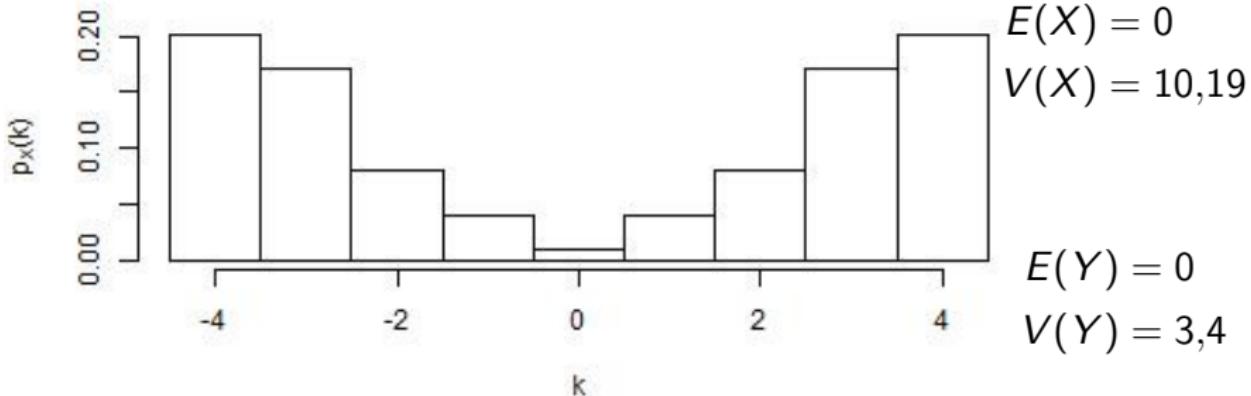


Calculamos

$$E[Y] = \sum_{k=-4}^4 kp_Y(k) = 0$$

$$E[Y^2] = \sum_{k=-4}^4 k^2 p_Y(k) = 3,4$$

$$V(Y) = E[Y^2] - (E[Y])^2 = 3,4$$



Esperanza y varianza de la distribución normal

Sea $Z \sim \mathcal{N}(0, 1)$. Queremos hallar $E(Z)$ y $V(Z)$.

$$\begin{aligned} E(|Z|) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |y| e^{-y^2/2} dy = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} ye^{-y^2/2} dy \\ &= \frac{1}{\sqrt{2\pi}} \left[-e^{-y^2/2} \right] \Big|_0^{\infty} = \frac{1}{\sqrt{2\pi}} [1] < +\infty \end{aligned}$$

Luego, la esperanza existe.

$$E(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ye^{-y^2/2} dy = 0$$

por ser la integral de una función integrable e impar, o sea,
 $h(y) = ye^{-y^2/2} = -h(-y)$.

Esperanza y varianza de la distribución normal

$$\begin{aligned} E(Z^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{[-y]}_u \underbrace{[-ye^{-y^2/2}]}_{v'} dy \\ &= \frac{1}{\sqrt{2\pi}} \left(\left[\underbrace{-y}_u \underbrace{e^{-y^2/2}}_v \right] \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \underbrace{(-1)}_{u'} \underbrace{e^{-y^2/2}}_v dy \right) \\ &= \underbrace{\frac{1}{\sqrt{2\pi}} \left[-ye^{-y^2/2} \right]_{-\infty}^{\infty}}_{=0} + \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy}_{=1} = 1 \end{aligned}$$

Luego,

$$E(Z) = 0 \quad V(Z) = 1$$

Esperanza y varianza de la distribución normal

Sea

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

Queremos hallar $E(X)$ y $V(X)$. Por el

Corolario 4.6 (estandarización de la normal) sabemos que

$Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$ tiene distribución normal estándar. Luego,

$X = \sigma Z + \mu$ y sabemos que $E(Z) = 0$ y $V(Z) = 1$. Finalmente, por linealidad de la esperanza tenemos,

$$E(X) = \sigma E(Z) + \mu = \mu,$$

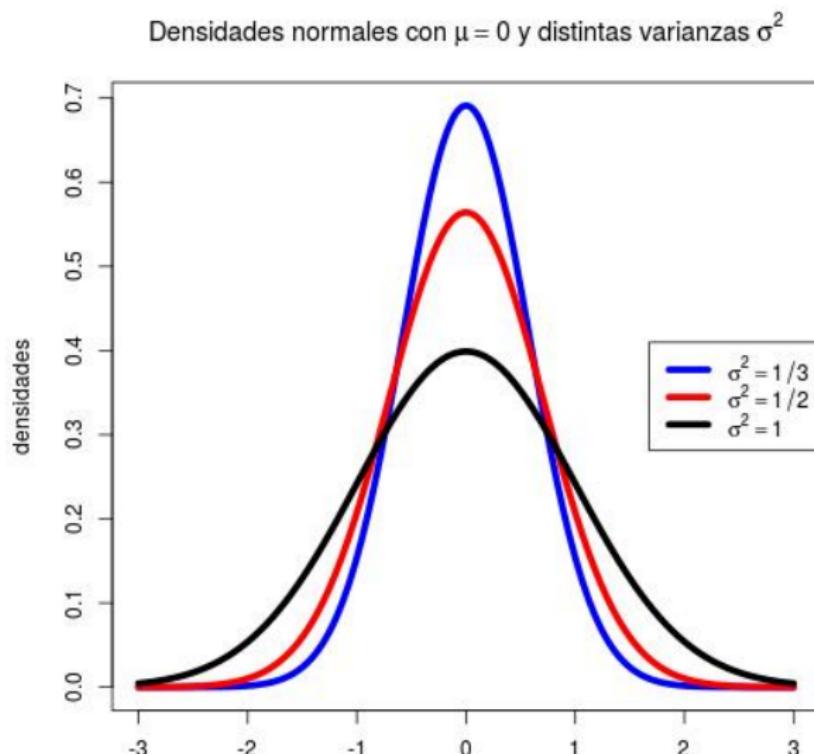
y por la **Propiedad V.3 (cambio de escala)** resulta

$$V(X) = \sigma^2 V(Z) = \sigma^2.$$

Es decir, que los dos parámetros con los que caracterizamos la distribución normal son su esperanza y su varianza.

Distribución Normal: varianza

a menor varianza, más concentrada



Espacio \mathcal{L}^2

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. Definimos el espacio

$\mathcal{L}^2(\Omega, \mathcal{F}, P)$ por

$$\mathcal{L}^2(\Omega, \mathcal{F}, P) = \{X : X \text{ es una variable aleatoria y } E[X^2] < +\infty\}.$$

Proposición 6.1

$\mathcal{L}^2(\Omega, \mathcal{F}, P)$ es un espacio vectorial con las operaciones

- $a \in \mathbb{R}$, $X \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$, entonces aX también está en $\mathcal{L}^2(\Omega, \mathcal{F}, P)$
- $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$, entonces $X + Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$.

Para probarlo necesitamos probar que se verifican todas las propiedades de espacios vectoriales, por ejemplo:

- a) $X = 0 \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$
- b) $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$, $a, b \in \mathbb{R}$, entonces $aX + bY \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$. O sea que sabiendo que $E[X^2] < +\infty$, $E[Y^2] < +\infty$ queremos probar que $E[(aX + bY)^2] < +\infty$.

Sabiendo que $E[X^2] < +\infty$, $E[Y^2] < +\infty$ qvq $E[(aX + bY)^2] < +\infty$. Sabemos que para todo $\alpha, \beta \in \mathbb{R}$, vale

$$(\alpha + \beta)^2 \leq 2\alpha^2 + 2\beta^2 \quad \text{¡comprobarlo!}$$

Luego $(aX + bY)^2 \leq 2a^2X^2 + 2b^2Y^2$ (desigualdad de funciones). Por la monotonía de la esperanza, y la linealidad, tenemos

$$E[(aX + bY)^2] \leq E[2a^2X^2 + 2b^2Y^2] = 2a^2E[X^2] + 2b^2E[Y^2] < +\infty$$

Espacio \mathcal{L}^2

Lema 6.10 (desigualdad de Cauchy – Schwarz)

$X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$, entonces $[E(XY)]^2 \leq E(X^2)E(Y^2)$.

La igualdad vale si y sólo si $P(aX = bY) = 1$ para $a, b \in \mathbb{R}$, al menos uno de los cuales es distinto de cero.

Demostración.

Si $E(Y^2) = 0$, como Y^2 es una variable positiva con esperanza 0, por el Lema 6.7 resulta que $Y^2 = 0$, luego $Y = 0$ y el lado izquierdo de la desigualdad que queremos probar vale 0 (el miembro derecho es siempre positivo). Más aun, vale la igualdad y tenemos $P(Y = 0) = 1$.

Si $E(Y^2) \neq 0$, sea

$g(t) = E([X + tY]^2) = E(X^2) + 2tE(XY) + t^2E(Y^2)$. g es una función cuadrática, $g : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$. Por ser positiva, **puede tener a lo sumo una raíz real**. Y su discriminante es ≤ 0 .

$\Delta = 4[E(XY)]^2 - 4E(Y^2)E(X^2) \leq 0$, y probamos la desigualdad. □

Espacio \mathcal{L}^2

Demostración.

¿Cuándo vale la igualdad? $[E(XY)]^2 - E(Y^2)E(X^2) = 0 \Leftrightarrow \Delta = 0$ si y sólo si la cuadrática tiene exactamente una raíz real, lo cual ocurre si y sólo si existe un t_0 tal que

$$g(t_0) = E([X + t_0 Y]^2) = 0$$

De nuevo, por el Lema 6.7, como $[X + t_0 Y]^2$ es una variable positiva que tiene esperanza cero, resulta que $[X + t_0 Y]^2 = 0$ con probabilidad 1, luego $P(X = -t_0 Y) = 1$. □

Observación 6.1

Si $X \in \mathcal{L}^2$, entonces $E[|X|] < +\infty$ pues, por la desigualdad de Cauchy – Schwarz tomando $Y = 1 \in \mathcal{L}^2$, tenemos

$$[E(|X|)]^2 = [E(|X|Y)]^2 \leq E(X^2)E(Y^2) = E(X^2)E(1) = E(X^2) < +\infty$$

Espacio \mathcal{L}^2 : producto interno

Proposición 6.2

$\langle X, Y \rangle = E[XY]$ define un producto interno en $\mathcal{L}^2(\Omega, \mathcal{F}, P)$.

La existencia de $\langle X, Y \rangle$ está garantizada por la desigualdad de Cauchy – Schwarz. Y además, resulta

- $\langle X, Y \rangle$ es bilineal, es decir,

$$E[(aX + bY)Z] = aE[XZ] + bE[YZ]$$

- $\langle X, Y \rangle$ es simétrico: trivial
- $\langle X, Y \rangle$ es definido positivo: sea X no nula (o sea, $P(X = 0) \neq 1$), entonces $\langle X, X \rangle > 0$.

Espacio \mathcal{L}^2 : norma

El producto interno define una norma en $\mathcal{L}^2(\Omega, \mathcal{F}, P)$,

$$\|X\|^2 = \langle X, X \rangle = E(X^2)$$

$$\text{ó } \|X\| = \sqrt{E(X^2)}.$$

y queda entonces definida la distancia

$$\|X - Y\| = \sqrt{E((X - Y)^2)}$$

La desigualdad triangular es consecuencia de la desigualdad de Cauchy – Schwarz, $\|X + Y\|^2 \leq (\|X\| + \|Y\|)^2$

$$\begin{aligned}\|X + Y\|^2 &= E[(X + Y)^2] = E[X^2] + E[Y^2] + 2E[XY] \\ &\leq E[X^2] + E[Y^2] + 2|E[XY]| \\ &\leq E[X^2] + E[Y^2] + 2\sqrt{E[X^2]}\sqrt{E[Y^2]} \\ &= \left(\sqrt{E[X^2]} + \sqrt{E[Y^2]}\right)^2 = (\|X\| + \|Y\|)^2\end{aligned}$$

Podemos formalizar lo hecho antes

Proposición 6.3

Sea $X \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$, entonces $\mu = E[X]$ es la constante que mejor approxima a X en el sentido que

$$E[(X - \mu)^2] \leq E[(X - c)^2], \quad \forall c \in \mathbb{R}$$

Demostración.

$$\begin{aligned} E[(X - c)^2] &= E[(X - \mu + \mu - c)^2] \\ &= E[(X - \mu)^2 + (\mu - c)^2 + 2(X - \mu)(\mu - c)] \\ &= E[(X - \mu)^2] + E[(\mu - c)^2] + E[2(X - \mu)(\mu - c)] \\ &= E[(X - \mu)^2] + (\mu - c)^2 + 2(\mu - c) E[X - \mu] \\ &= E[(X - \mu)^2] + (\mu - c)^2 \geq E[(X - \mu)^2] \quad \text{si } c \neq \mu. \end{aligned}$$



Covarianza

Definición 6.5 (covarianza)

Dadas dos variables aleatorias X e $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$, se define la **covarianza** entre ellas mediante la fórmula

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] ,$$

siendo $\mu_X = E[X]$ y $\mu_Y = E[Y]$.

Lema 6.11 (Fórmula reducida para la covarianza)

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] .$$

Demostración.

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y]$$

Covarianza: propiedades

Teorema 6.12

Si X e Y son independientes, entonces

$$\text{Cov}(X, Y) = 0$$

Demostración.

Por el Lema 6.8 (esperanza de producto de variables independientes), tenemos que

$$\begin{aligned}\text{Cov}(X, Y) &= E \left[\underbrace{(X - \mu_X)}_{g(X)} \underbrace{(Y - \mu_Y)}_{h(Y)} \right] \\ &= E[g(X)] E[h(Y)] = \underbrace{E[X - \mu_X]}_{=0} \underbrace{E[Y - \mu_Y]}_{=0} = 0\end{aligned}$$



Covarianza

¡Cuidado! Covarianza cero ¡NO garantiza independencia!

Ejercicio 6.1

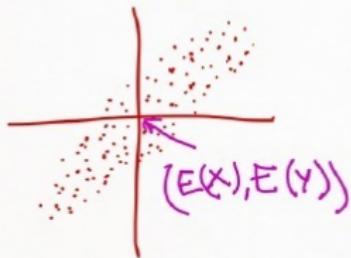
Sea $X \sim \mathcal{U}(-1, 1)$. Sea $Y = X^2$. Demuestre que $\text{Cov}(X, Y) = 0$. ¿Son las variables X e Y independientes?

Sugerencia: calcule $P(X > 1/2, Y < 1/4)$.

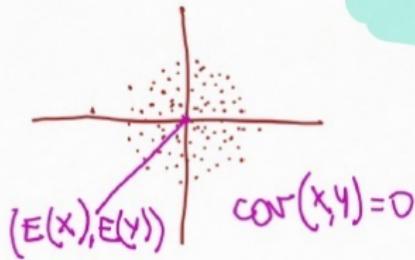
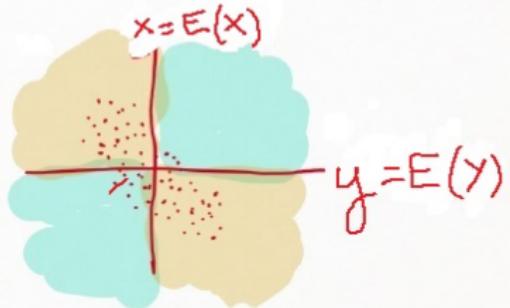
Covarianza: interpretación del signo

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$\text{cor}(x, y) > 0$$



$$\text{cor}(x, y) < 0$$



$$\text{cor}(x, y) = 0$$

Covarianza

Lema 6.13

La covarianza verifica las siguientes propiedades:

- ① $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- ② $\text{Cov}(X, X) = V(X)$
- ③ $\text{Cov}(a + X, Y) = \text{Cov}(X, Y), \quad a \in \mathbb{R}$
- ④ $\text{Cov}(aX, bY) = a b \text{Cov}(X, Y), \quad a, b \in \mathbb{R}$
- ⑤ $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$
- ⑥ $\text{Cov}(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(X_i, Y_j), \text{ con } a_i, b_j \in \mathbb{R}$
- ⑦ $V(X + Y) = V(X) + V(Y) + 2 \text{Cov}(X, Y)$

3)

$$\begin{aligned}\text{Cov}(a + X, Y) &= E([a + X - E(a + X)][Y - E(Y)]) \\ &= E([X - E(X)][Y - E(Y)]) = \text{Cov}(X, Y)\end{aligned}$$

4) Como $E(ax) = aE(X)$,

$$\begin{aligned}\text{Cov}(aX, bY) &= E([aX - aE(X)][bY - bE(Y)]) \\ &= E(ab[X - E(X)][Y - E(Y)]) \\ &= abE([X - E(X)][Y - E(Y)]) \\ &= ab \text{Cov}(X, Y)\end{aligned}$$

5) Consideremos $\text{Cov}(X, Y + Z)$

$$\begin{aligned}\text{Cov}(X, Y + Z) &= E([X - E(X)]\{[Y - E(Y)] + [Z - E(Z)]\}) \\ &= E([X - E(X)][Y - E(Y)] + [X - E(X)][Z - E(Z)]) \\ &= E([X - E(X)][Y - E(Y)]) \\ &\quad + E([X - E(X)][Z - E(Z)]) = \text{Cov}(X, Y) + \text{Cov}(X, Z)\end{aligned}$$

6) Lo obtenemos combinando 1), 4) y 5) .

Covarianza: propiedades

Faltaba probar

$$7) V(X + Y) = V(X) + V(Y) + 2 \operatorname{Cov}(X, Y)$$

Demostración.

$$\begin{aligned} V(X + Y) &= \operatorname{Cov}(X + Y, X + Y) = \operatorname{Cov}(X, X) + \operatorname{Cov}(X, Y) + \\ &\quad \operatorname{Cov}(Y, X) + \operatorname{Cov}(Y, Y) = V(X) + V(Y) + 2 \operatorname{Cov}(X, Y) \end{aligned}$$



Ejercicio 6.2

Escribir la fórmula de $V(X - Y)$.

Ejercicio 6.3

Usar la propiedad 7 para calcular la varianza de una $X \sim \mathcal{B}(n, p)$

Covarianza: propiedades

Lema 6.14 (Cauchy–Schwarz para covarianza)

$$|\text{Cov}(Z, W)|^2 \leq V(Z)V(W).$$

La igualdad vale si y sólo si $P(a_1 Z = a_2 W + a_3) = 1$ para $a_1, a_2, a_3 \in \mathbb{R}$ con $a_1 \neq 0$ ó $a_2 \neq 0$.

Demostración.

Usamos el Lema 6.10, con $X = Z - \mu_Z$ e $Y = W - \mu_W$,

$$\begin{aligned} |\text{Cov}(Z, W)|^2 &= |E[(Z - \mu_Z)(W - \mu_W)]|^2 = [E(XY)]^2 \\ &\leq E(X^2) E(Y^2) = E[(Z - \mu_Z)^2] E[(W - \mu_W)^2] \\ &= V(Z)V(W). \end{aligned}$$

El Lema 6.10 afirma que la igualdad vale si y sólo si $P(aX = bY) = 1$ para $a, b \in \mathbb{R}$, al menos uno de los cuales es distinto de cero. O sea $a(Z - \mu_Z) = b(W - \mu_W)$, con lo cual la afirmación vale. □

Covarianza de la normal bivariada

Un vector aleatorio (X, Y) tiene distribución normal bivariada si su función de densidad conjunta es

$$f_{XY}(x, y) = \frac{1}{C} e^{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) \right]}.$$

con $C = 2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}$. Calculemos la $\text{Cov}(X, Y)$.

$$\text{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{XY}(x, y) dx dy$$

Haciendo el cambio de variables $u = (x - \mu_X)/\sigma_X$ y $v = (y - \mu_Y)/\sigma_Y$ con Jacobiano $\sigma_X\sigma_Y$

$$\frac{\cdot \sigma_X\sigma_Y}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} uv \exp \left[-\frac{1}{2(1-\rho^2)} (u^2 + v^2 - 2\rho uv) \right] du dv$$

Covarianza de la normal bivariada

Para calcular la integral, completamos cuadrados en u . Usamos la igualdad:

$$u^2 + v^2 - 2\rho uv = (u - \rho v)^2 + v^2 (1 - \rho^2)$$

Entonces, podemos reemplazar en la integral que teníamos

$$\frac{\sigma_X \sigma_Y}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} uv \exp \left[-\frac{1}{2(1 - \rho^2)} (u^2 + v^2 - 2\rho uv) \right] du dv$$

y obtener

$$\frac{\sigma_X \sigma_Y}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{\infty} v \exp(-v^2/2) \left(\int_{-\infty}^{\infty} u \exp \left[-\frac{1}{2(1 - \rho^2)} (u - \rho v)^2 \right] du \right) dv$$

La **integral interior** es la esperanza de una v.a. $\mathcal{N}(\rho v, (1 - \rho^2))$, a la que únicamente le falta la constante de normalización $[2\pi (1 - \rho^2)]^{-1/2}$, y por lo tanto tenemos

$$\text{Cov}(X, Y) = \rho \sigma_X \sigma_Y \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} v^2 e^{-v^2/2} dv = \rho \sigma_X \sigma_Y$$

(la expresión azul es la varianza de la normal estándar, que es 1)

Correlación

Definición 6.6 (coeficiente de correlación)

*Dadas dos variables aleatorias X e Y definimos el **coeficiente de correlación** entre ellas mediante la fórmula*

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}}.$$

Decimos que X e Y son **no correlacionadas** cuando $\text{Cov}(X, Y) = 0$. La covarianza no es por sí misma una medida muy satisfactoria de la dependencia entre X e Y , pues el valor que toma se ve afectado por cambios de la escala de las variables, $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$. El siguiente Lema muestra que eso no pasa con la correlación, lo cual la hace más interpretable.

Propiedades de la Correlación

Lema 6.15

$$-1 \leq \rho(X, Y) \leq 1$$

Además, $|\rho(X, Y)| = 1$ si y sólo si existen constantes $a, b \in \mathbb{R}$ para las cuales $P(Y = aX + b) = 1$.

Demostración.

Es una consecuencia del Lema 6.14. □

Lema 6.16

$$\rho(aX, bY) = \rho(X, Y), \quad \forall a, b \in \mathbb{R}.$$

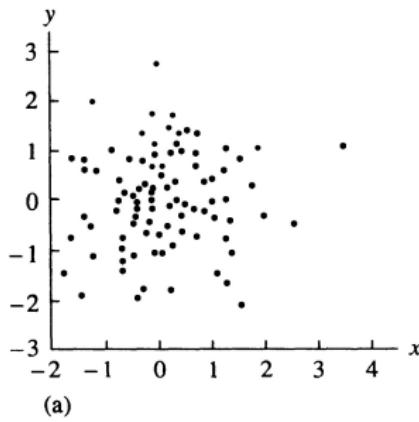
Ejercicio 6.4

Probar el Lema 6.16.

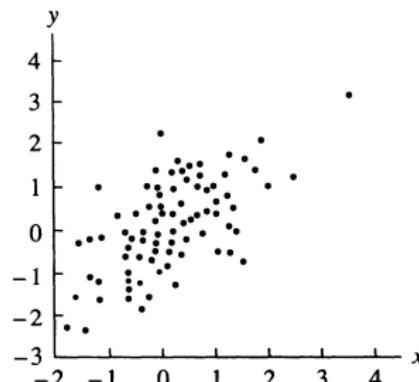
Ejercicio 6.5

Sea $(X, Y) \sim \mathcal{N}_2 \left((\mu_X, \mu_Y), \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right)$, un vector aleatorio normal bivariado. Basándonos en las cuentas que ya hicimos, deducir el valor de $\rho(X, Y)$.

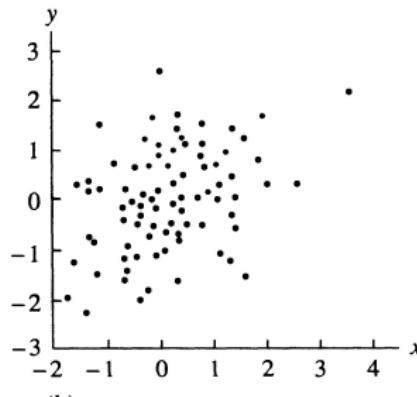
El coeficiente de correlación $\rho(X, Y)$ mide la fuerza de la relación lineal. entre X e Y (compárese con los gráficos de las curvas de nivel que hicimos en el capítulo de vectores aleatorios). La correlación también afecta la apariencia de un diagrama de dispersión, o *scatterplot* que se construye generando n pares independientes (X_i, Y_i) , donde $i = 1, \dots, n$ y luego graficándolos. La figura que sigue muestra diagramas de dispersión de 100 pares de observaciones de variables aleatorias normales bivariadas pseudoaleatorias para varios valores de ρ . Observemos que las nubes de puntos tienen una forma aproximadamente elíptica.



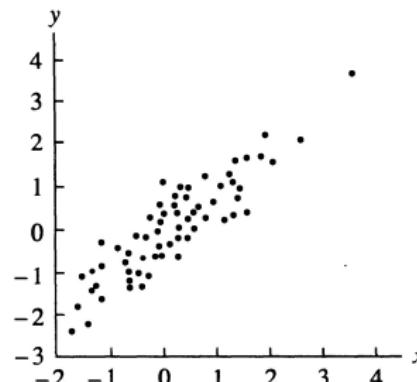
(a)



(c)



(b)



(d)

- (a) $\rho = 0$
- (b) $\rho = 0,3$
- (c) $\rho = 0,6$
- (d) $\rho = 0,9$

7. Desigualdades y Convergencia

Probabilidades y Estadística (M)

María Eugenia Szretter Noste

Departamento de Matemática e
Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Primer cuatrimestre 2020



7. Desigualdades que relacionan probabilidad y esperanza

Proposición 7.1 (Desigualdad de Markov)

Sea $X \geq 0$, $E(X) < +\infty$, $\lambda > 0$. Entonces

$$P(X \geq \lambda) \leq \frac{E(X)}{\lambda}$$

Sea $Y = \lambda I_{[\lambda, +\infty)}(X) = \lambda I\{X \geq \lambda\}$.
Entonces $Y \leq X$ (observar que Y es discreta)

7. Desigualdades que relacionan probabilidad y esperanza

Proposición 7.1 (Desigualdad de Markov)

Sea $X \geq 0$, $E(X) < +\infty$, $\lambda > 0$. Entonces

$$P(X \geq \lambda) \leq \frac{E(X)}{\lambda}$$

Sea $Y = \lambda I_{[\lambda, +\infty)}(X) = \lambda I\{X \geq \lambda\}$.

Entonces $Y \leq X$ (observar que Y es discreta)
Por la monotonía de la esperanza,

7. Desigualdades que relacionan probabilidad y esperanza

Proposición 7.1 (Desigualdad de Markov)

Sea $X \geq 0$, $E(X) < +\infty$, $\lambda > 0$. Entonces

$$P(X \geq \lambda) \leq \frac{E(X)}{\lambda}$$

Sea $Y = \lambda I_{[\lambda, +\infty)}(X) = \lambda I\{X \geq \lambda\}$.

Entonces $Y \leq X$ (observar que Y es discreta)
Por la **monotonía** de la esperanza,

$$E(Y) \leq E(X)$$

7. Desigualdades que relacionan probabilidad y esperanza

Proposición 7.1 (Desigualdad de Markov)

Sea $X \geq 0$, $E(X) < +\infty$, $\lambda > 0$. Entonces

$$P(X \geq \lambda) \leq \frac{E(X)}{\lambda}$$

Sea $Y = \lambda I_{[\lambda, +\infty)}(X) = \lambda I\{X \geq \lambda\}$.

Entonces $Y \leq X$ (observar que Y es discreta)

Por la **monotonía** de la esperanza,

$$E(Y) \leq E(X)$$

$$\lambda P(X \geq \lambda) \leq E(X) \quad \checkmark$$

Desigualdades que relacionan probabilidad y esperanza

Corolario 7.1

Sea $\varphi : \mathbb{R} \rightarrow [0, \infty)$ creciente (en sentido amplio), y sea X una variable aleatoria tal que $E(\varphi(X)) < \infty$. Entonces, para todo $\lambda > 0$,

$$P(X \geq \lambda) \leq \frac{E(\varphi(X))}{\varphi(\lambda)}$$

dem: $x \geq \lambda \Rightarrow \varphi(x) \geq \varphi(\lambda)$ por ser φ creciente

$$\{x \geq \lambda\} \subset \{\varphi(x) \geq \varphi(\lambda)\}$$

por monotonía $\Rightarrow P(x \geq \lambda) \leq P(\varphi(x) \geq \varphi(\lambda)) \stackrel{\text{Markov}}{\leq} \frac{E(\varphi(x))}{\varphi(\lambda)}$

Desigualdades que relacionan probabilidad y esperanza

Proposición 7.2 (Desigualdad de Chebychev)

Sea X una variable aleatoria con $E(X) < +\infty$, $V(X) < +\infty$. Para todo $\varepsilon > 0$,

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}$$

dem: Tomemos la variable $y = |X - E(x)| \geq 0$

y la función $\varphi(t) = t^2$ que es creciente en $[0, +\infty)$
 $\varphi: [0, +\infty) \rightarrow [0, +\infty)$

$$\begin{aligned} P(|X - E(x)| \geq \varepsilon) &= P(y \geq \varepsilon) \leq \frac{E(y^2)}{\varepsilon^2} = \frac{E(|X - E(x)|^2)}{\varepsilon^2} \\ &= \frac{V(X)}{\varepsilon^2} \quad \checkmark \end{aligned}$$


cordario anterior

Aplicaciones

La desigualdad de Chebycher le da sustento a la afirmación de que el desvío estándar es una medida de dispersión de una variable.

Sea $\mu = E(X)$ y $\sigma^2 = V(X)$

Desig de Chebycher: $P(|X - \mu| \geq \epsilon) \leq \frac{V(X)}{\epsilon^2}$

$$\Leftrightarrow P(|X - \mu| < \epsilon) \geq 1 - \frac{V(X)}{\epsilon^2}$$

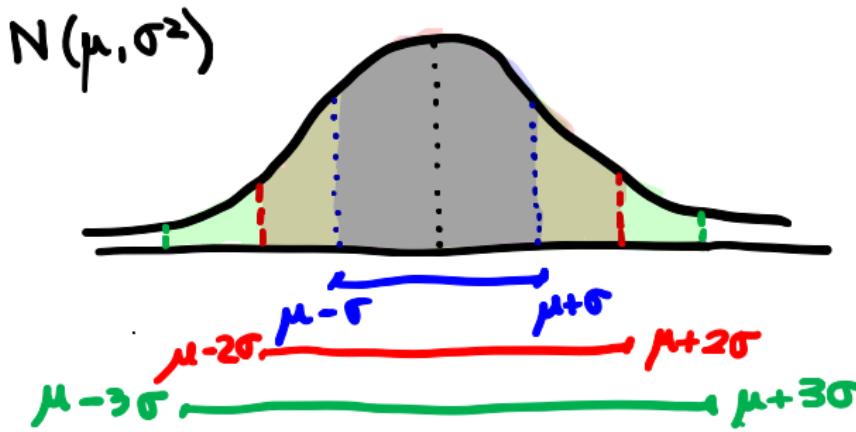
Tomando $\epsilon = k\sigma$

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

$$\Leftrightarrow P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - 1/k^2$$

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - 1/k^2$$

k	intervalo	cota	$N(\mu, \sigma^2)$
1	$(\mu - \sigma, \mu + \sigma)$	0	0,6826
2	$(\mu - 2\sigma, \mu + 2\sigma)$	$\frac{3}{4} = 0,75$	0,9544
3	$(\mu - 3\sigma, \mu + 3\sigma)$	$\frac{8}{9} = 0,89$	0,9977



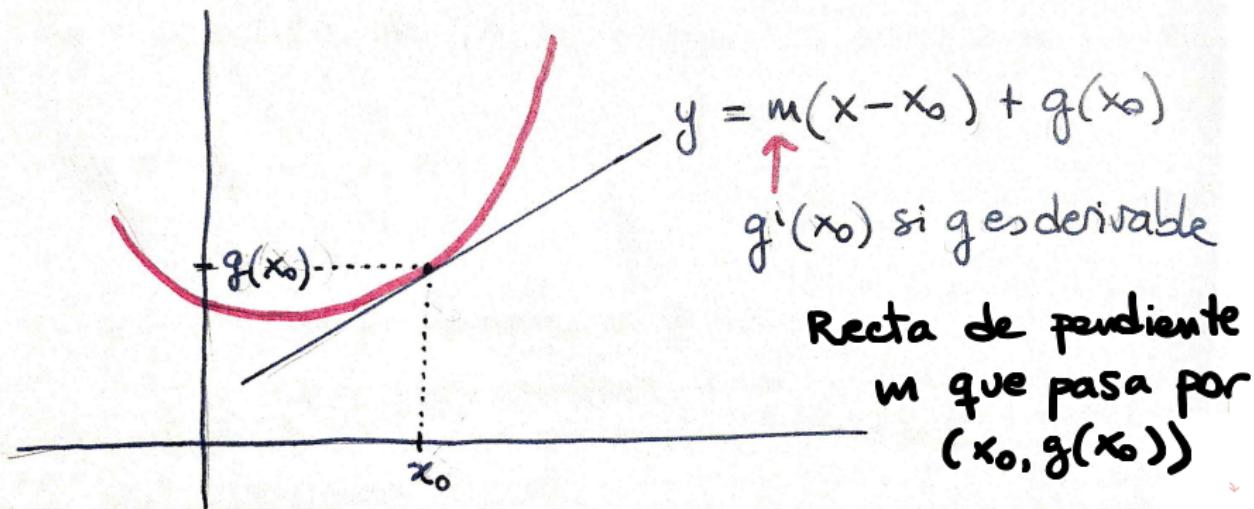
- la cota se cumple
- es mala pero universal
- fijada la probab, a mayor σ , + ancho el intervalo

Desigualdades que relacionan probabilidad y esperanza

Definición 7.1 (función convexa)

Decimos que una función $g : \mathbb{R} \rightarrow \mathbb{R}$ es **convexa** si para cada $x_0 \in \mathbb{R}$ existe una constante m (en los casos buenos, $m = g'(x_0)$) tal que $\forall x \in \mathbb{R}$,

$$g(x) \geq g(x_0) + m(x - x_0)$$



Desigualdades

Proposición 7.3 (Desigualdad de Jensen)

Sea $g : \mathbb{R} \rightarrow \mathbb{R}$ convexa y sea X una variable aleatoria con $E|X| < +\infty$, $E|g(X)| < +\infty$. Entonces,

$$g(E(X)) \leq E(g(X)).$$

dem: Tomamos $x_0 = E(x)$ en la definición de convexidad

Sea m/ $g(t) \geq g(E(x)) + m(t - E(x)) \quad \forall t \in \mathbb{R}$

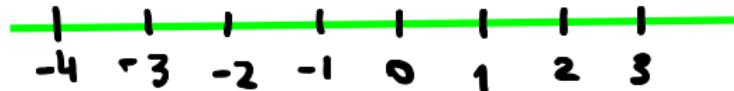
Luego $g(x) \geq g(E(x)) + m(x - E(x))$

Tomando esperanza, x monotonía tenemos:

$$E(g(x)) \geq g(E(x)) + m \underbrace{E(x - E(x))}_{=0} = g(E(x))$$

Paseo al azar (simple random walk)

Una partícula se mueve ocupando los lugares enteros de la recta. En cada paso se mueve un



paso a la derecha (con probabilidad p) o a la izquierda (con probabilidad $1-p$). La dirección que elige en cada paso (der. o izq.) que elige en cada paso es independiente de las demás.

Sea

s_n = posición de la partícula a tiempo n

$$s_n = s_0 + \sum_{i=1}^n x_i \quad \text{con}$$

$$x_i = \begin{cases} 1 & \text{con probab } p \\ -1 & \text{con probab } 1-p \end{cases}$$

s_0 es la posición inicial

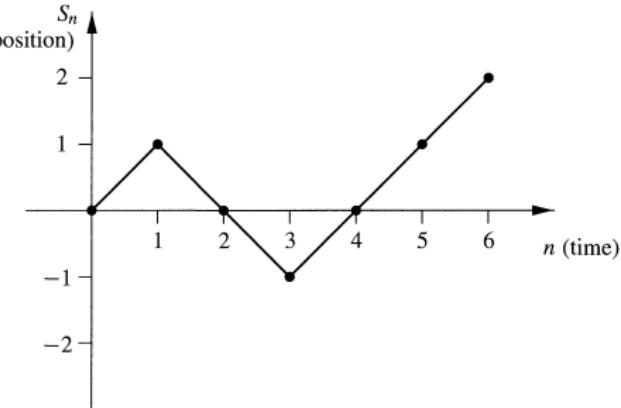
$$S_n = S_0 + \sum_{i=1}^n X_i \quad \text{con} \quad X_i = \begin{cases} 1 & \text{con probab } p \\ -1 & \text{con probab } 1-p \end{cases}$$

$\{X_i\}_{i \geq 1}$ son variables aleatorias independientes

Un reescalamiento de cada X_i tiene distribución $Be(p)$:

$$Y_i = \frac{1}{2}X_i + \frac{1}{2} \sim Be(p)$$

Registraremos el movimiento de la partícula a través de la sucesión $\{(n, S_n) : n \geq 0\}$ de puntos del plano. Esta colección de puntos, unidos por segmentos entre vecinos, se llama el paseo de la partícula.



← Un ejemplo de un paseo al azar. La partícula caminó y fue visitando puntos $0, 1, 0, -1, 0, 1, 2$ en sucesión. $S_0 = 0$.

Estamos interesados en el evento

$\{S_k=0\} = \{\text{la partícula retornó al origen a tiempo } n\}$

Sólo es posible si $k=2n$, entonces

$$P(S_k=0)=0 \text{ si } k=2n-1, \forall n \in \mathbb{N}$$

$$P(S_{2n} = 0) = P\left(S_0 + \sum_{i=1}^{2n} X_i = 0\right) = P\left(\sum_{i=1}^{2n} (2Y_i - 1) = 0\right)$$

.

$$Y_i \sim Be(p), X_i = 2Y_i - 1$$

$$S_0 = 0$$

$$= P\left(2\left[\sum_{i=1}^{2n} Y_i\right] - 2n = 0\right) = P\left(\sum_{i=1}^{2n} Y_i = n\right)$$

$$Y_i \sim Be(p) \text{ indep} \Rightarrow \sum_{i=1}^{2n} Y_i \sim Bi(2n, p)$$

$$\stackrel{\curvearrowleft}{=} \binom{2n}{n} p^n (1-p)^n.$$

$$\rightarrow P(S_k = 0) = \begin{cases} 0 & \text{si } k \text{ es impar} \\ \binom{2n}{n} [p(1-p)]^n & \text{si } k = 2n \\ & (n \in \mathbb{N}) \end{cases}$$

Consideremos $\Omega = \{ \omega = (w_1, w_2, \dots) : w_i \in \{-1, 1\} \}$

las sucesiones de ± 1 . Tenemos definidas las r.a.

$X_i : \Omega \rightarrow \mathbb{R}$, $S_n : \Omega \rightarrow \mathbb{R}$ dadas por

$$X_i(\omega) = w_i$$

$S_n(\omega) = \sum_{i=1}^n X_i(\omega)$. Nos interesa una variable

aleatoria $Y : \Omega \rightarrow \mathbb{N}$ dada por

$Y(\omega) = \text{Última visita al Origen de la trayectoria } \omega$

$$Y(\omega) = \max \{ n \in \mathbb{N} : S_n = 0 \}$$

¿Está bien definida esta variable aleatoria? Es decir,
¿existe una última visita al origen para cada paseo al azar?

Sean $A_n = \{ S_n = 0 \} \quad n \geq 1$

Límites Inferiores y Superiores

Dada una sucesión de eventos $(A_n)_{n \geq 1}$, consideremos dos nuevos eventos.

$$\begin{aligned} A^\infty &= \limsup_{n \rightarrow \infty} A_n = \bigcap_{k \geq 1} \bigcup_{n \geq k} A_n \\ &= \{\omega \in \Omega : \omega \in A_n \text{ para infinitos valores de } n\} \end{aligned}$$

También se lo escribe $\{A_n \text{ ocurre infinitas veces}\}$, o, $\{A_n \text{ i.o.}\}$, (i.o. son las siglas de *infinitely often*).

$$w \in \bigcap_{k \geq 1} \bigcup_{n \geq k} A_n \Leftrightarrow \forall k \in \mathbb{N}, w \in \bigcup_{n \geq k} A_n \Leftrightarrow w \in A_n \text{ para infinitos } n \in \mathbb{N}$$

$$\begin{aligned} A_\infty &= \liminf_{n \rightarrow \infty} A_n = \bigcup_{k \geq 1} \bigcap_{n \geq k} A_n \\ &= \{\omega \in \Omega : \omega \in A_n \text{ para todos los } A_n \text{ excepto para un número finito de ellos}\} \end{aligned}$$

$$w \in \bigcup_{k=1}^{\infty} \bigcap_{n \geq k} A_n \Leftrightarrow \exists k \in \mathbb{N} : w \in \bigcap_{n \geq k} A_n \Leftrightarrow w \in A_n \quad \forall n \geq k$$

(k = k(w))

Paseo al azar

En el ejemplo del paseo al azar, ¿qué sería el conjunto $\{A_n \text{ ocurre infinitas veces}\}$?

Sería el conjunto de todas las trayectorias o paseos de la partícula que la lleva a volver cada tanto al origen y para las cuales no queda definida la v.a. Y última visita al origen.

$$\{A_n \text{ ocurre infinitas veces}\} = \limsup_{n \rightarrow \infty} A_n \text{ con } A_n = \{S_n = 0\}$$

Nos interesa calcular $P(A_n \text{ ocurre infinitas veces})$. Miremos algunas propiedades de los límites sup cinf y leva de Borell-Cantelli

Límites Inferiores y Superiores: Propiedades

Lema 7.2

- 1) Sea $B_k = \bigcup_{n \geq k} A_n$, entonces la sucesión $(B_k)_{k \geq 1}$ es una sucesión decreciente de eventos, o sea $B_{k+1} \subset B_k$, $\forall k \in \mathbb{N}$, luego

$$\limsup_{n \rightarrow \infty} A_n = \bigcap_{k \geq 1} B_k, \text{ o sea } B_k \searrow \limsup_{n \rightarrow \infty} A_n$$

y por lo tanto, $\lim_{k \rightarrow \infty} P(B_k) = P(\limsup_{n \rightarrow \infty} A_n)$

- 2) Sea $C_k = \bigcap_{n \geq k} A_n$, entonces la sucesión $(C_k)_{k \geq 1}$ es una sucesión creciente de eventos, o sea $C_k \subset C_{k+1}$, $\forall k \in \mathbb{N}$, luego

$$\liminf_{n \rightarrow \infty} A_n = \bigcup_{k \geq 1} C_k, \text{ o sea } C_k \nearrow \liminf_{n \rightarrow \infty} A_n$$

y por lo tanto, $\lim_{k \rightarrow \infty} P(C_k) = P(\liminf_{n \rightarrow \infty} A_n)$

Lema 7.2

3)

$$\left(\limsup_{n \rightarrow \infty} A_n \right)^c = \left(\bigcap_{k \geq 1} \bigcup_{n \geq k} A_n \right)^c = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n^c = \liminf_{n \rightarrow \infty} (A_n^c)$$

4) $I_{\limsup A_n} = \limsup_{n \rightarrow \infty} I_{A_n}$, $I_{\liminf A_n} = \liminf_{n \rightarrow \infty} I_{A_n}$

5) $\liminf_{n \rightarrow \infty} A_n \subset \limsup_{n \rightarrow \infty} A_n$

6) $C_k = \underbrace{\bigcap_{n \geq k} A_n}_{suc. creciente} \subset A_k \subset \underbrace{\bigcup_{n \geq k} A_n}_{suc. decreciente} = B_k$

Lema 7.3 (Borell Cantelli)

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. Considere una sucesión $(A_n)_{n \geq 1}$ de eventos: $A_n \in \mathcal{F}$ para todo n .

- i. Si $\sum_{n \geq 1} P(A_n) < \infty$, entonces $P\left(\underbrace{\bigcap_{k \geq 1} \bigcup_{n \geq k} A_n}_{\limsup_n A_n}\right) = 0$.
- ii. Si los eventos $(A_n)_{n \geq 1}$ son independientes y $\sum_{n \geq 1} P(A_n) = +\infty$, entonces $P\left(\bigcap_{k \geq 1} \bigcup_{n \geq k} A_n\right) = 1$.

Demostración.

i.

$$P\left(\bigcap_{k \geq 1} \bigcup_{n \geq k} A_n\right) \leq P\left(\bigcup_{n \geq k} A_n\right) \leq \sum_{n \geq k} P(A_n) \xrightarrow{k \rightarrow \infty} 0$$

por ser la cola de una serie convergente.

Demostración.

ii.

$$P\left(\bigcap_{k \geq 1} \bigcup_{n \geq k} A_n\right) = \lim_{k \rightarrow \infty} P\left(\bigcup_{n \geq k} A_n\right) = \lim_{k \rightarrow \infty} \left[1 - P\left(\bigcap_{n \geq k} A_n^c\right)\right]$$

Pero

$$\begin{aligned} P\left(\bigcap_{n \geq k} A_n^c\right) &= \lim_{N \rightarrow \infty} P\left(\bigcap_{n=k}^N A_n^c\right) \underset{\text{por indep}}{=} \lim_{N \rightarrow \infty} \prod_{n=k}^N P(A_n^c) \\ &= \lim_{N \rightarrow \infty} \prod_{n=k}^N (1 - P(A_n)) \leq \lim_{N \rightarrow \infty} \prod_{n=k}^N e^{-P(A_n)} \quad (1 - x \leq e^{-x} \text{ si } x \geq 0) \\ &= \lim_{N \rightarrow \infty} e^{-\sum_{n=k}^N P(A_n)} = 0 \end{aligned}$$



La independencia es necesaria para que valga (ii) de Borel –Cantelli.

Ejemplo 7.1

Sea E un evento con $0 < P(E) < 1$. Sea $A_n = E \forall n \in \mathbb{N}$.

Entonces,

- los $(A_n)_{n \geq 1}$ no son independientes entre sí.
- $\sum_{n \geq 1} P(A_n) = +\infty$,
- $\limsup_n A_n = E$ y $P(\limsup_n A_n) = P(E) < 1$.

Repetimos un mismo experimento de forma independiente. Nos interesa un evento con probabilidad positiva (que podría ser muy pequeña) y nos preguntamos si ese resultado ocurrirá alguna vez, Borell– Cantelli (ii) nos dice que no sólo ocurrirá una vez, sino que la probabilidad de que ocurra infinitas veces ¡es uno!

Paseo al azar

Para calcular $P(A_n \text{ ocurre infinitas veces})$ podemos usar Borel - Cantelli. Para eso calcularemos:

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} P(S_n = 0) = \sum_{k=1}^{\infty} \binom{2k}{k} p^k (1-p)^k < \infty$$

↑ si $p \neq 1/2$

\Rightarrow $(\times \text{BCI})$ $P(A_n \text{ inf veces}) = 0$ $\xrightarrow{\text{si } p \neq 1/2}$
Veamos que la serie converge. Por inducción es fácil ver que $\binom{2k}{k} \leq 4^k \quad \forall k \in \mathbb{N}$.

Luego $\binom{2k}{k} p^k (1-p)^k \leq \underbrace{[4p(1-p)]^k}_{a_k}$ $\xrightarrow{\text{c para qué valores de } p \text{ es }} a_k < 1 ?$

$$a_k = [4p(1-p)]^k < 1 \Leftrightarrow 4p(1-p) < 1$$

Pero $g(p) = 4p(1-p)$ es una cuadrática con raíces en 0 y 1, que alcanza su máximo en $p = 1/2$.

Luego $g(p) \leq g(1/2) = 4 \cdot \frac{1}{4} = 1$.

$g(p) < 1 \quad \forall p \neq 1/2$, por lo que $\sum_{n=1}^{\infty} P(A_n) < \infty$

Usando la fórmula de Stirling para aproximar, si $p \neq 1/2$
 puede verse que $\sum_{n=1}^{\infty} P(A_n) = \infty$ cuando $\underbrace{p}_{\uparrow} = 1/2$. ¿Podemos

dicir algo de $P(\limsup_{n \rightarrow \infty} A_n)$ vía Borell-Cantelli? No, xq
 los eventos $(A_n)_{n \geq 1}$ en este caso no son independientes.

Convergencia casi segura o en casi todo punto

Definición 7.2 (Convergencia en casi todo punto)

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. $X_n, X : \Omega \rightarrow \mathbb{R}$ variables aleatorias. Diremos que la sucesión de variables aleatorias $(X_n)_{n \geq 1}$ **converge en casi todo punto** a la variable aleatoria X , y lo anotaremos $X_n \xrightarrow{\text{c.t.p.}} X$ o $X_n \xrightarrow{\text{c.s.}} X$ si $\{\omega : X_n(\omega) \rightarrow X(\omega)\}$ es un evento con probabilidad 1.

Una caracterización sumamente útil del conjunto de puntos del espacio muestral donde Y_n es convergente está dada por la identidad

$$C = \{\omega : X_n(\omega) \rightarrow X(\omega)\} = \bigcap_{M > 0} \bigcup_{n_0 > 0} \bigcap_{n \geq n_0} \{|X_n - X| \leq 1/M\} .$$

Notemos que elegimos trabajar con $1/M$ para cantidades chicas de forma tal de garantizar que estamos siempre operando numerablemente, lo que garantiza que no nos salimos de la σ -álgebra \mathcal{F} , o sea, $C \in \mathcal{F}$.

(copiamos)

$$C = \{\omega : X_n(\omega) \rightarrow X(\omega)\} = \bigcap_{M \geq 1} \bigcup_{n_0 \geq 1} \bigcap_{n \geq n_0} \underbrace{\{ |X_n - X| \leq 1/M \}}_{\liminf \{ |X_n - X| \leq 1/M \}}$$

$\omega \in C \Leftrightarrow \forall \varepsilon > 0 \exists n_0 : \forall n \geq n_0 = n_0(\omega) \text{ vale que } |X_n(\omega) - X(\omega)| < \varepsilon$
 $\Leftrightarrow \forall M > 0 \exists n_0 = n_0(\omega) : \forall n \geq n_0 \text{ vale que } |X_n(\omega) - X(\omega)| < \frac{1}{M}$

Tomando complementos,

$$\begin{aligned} C^c &= \{\omega : X_n(\omega) \not\rightarrow X(\omega)\} = \bigcup_{M \geq 1} \bigcap_{n_0 \geq 1} \bigcup_{n \geq n_0} \underbrace{\{ |X_n - X| > 1/M \}}_{A_n(1/M)} \\ &= \bigcup_{M \geq 1} \limsup_{n \rightarrow \infty} A_n(1/M) \end{aligned}$$

Tenemos entonces que $X_n \xrightarrow{c.t.p.} X$ si y sólo si para todo M vale que

$$P \left(\bigcap_{n_0} \bigcup_{n \geq n_0} \{ |X_n - X| > 1/M \} \right) = 0.$$

En realidad, vale el siguiente resultado:

Lema 7.4

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. $X_n, X : \Omega \rightarrow \mathbb{R}$ variables aleatorias. Son equivalentes

a) $X_n \xrightarrow{\text{c.t.p.}} X$

b) Para todo $\varepsilon > 0$ vale que

$$P \left(\bigcap_{n_0} \bigcup_{n \geq n_0} \{|X_n - X| > \varepsilon\} \right) = 0.$$

c) Para todo $\varepsilon > 0$ vale que

$$\lim_{n_0 \rightarrow \infty} P \left(\bigcup_{n \geq n_0} \{|X_n - X| > \varepsilon\} \right) = 0.$$

Dem:

a) \Rightarrow b) $x_n \xrightarrow{\text{ctp}} x \Rightarrow P(C^c) = 0$ siendo

$$C^c = \{w \in \Omega : x_n(w) \not\rightarrow x(w)\} . \text{ Sea } \varepsilon > 0$$

Entonces

$$\bigcap_{k=1}^{\infty} \bigcup_{n \geq k} \underbrace{\{ |x_n - x| > \varepsilon \}}_{A_n(\varepsilon)} = \limsup_n A_n(\varepsilon) \subset C^c$$

pues si $w \in \bigcap_{k=1}^{\infty} \bigcup_{n \geq k} \{ |x_n - x| > \varepsilon \}$ $\Rightarrow \forall k \in \mathbb{N}, \exists n \geq k / |x_n(w) - x(w)| > \varepsilon$
 $\therefore x_n(w) \not\rightarrow x(w) \Rightarrow w \in C^c$.

Luego $P(\limsup_n A_n(\varepsilon)) \leq P(C^c) = 0$ ✓

b) \Leftrightarrow c) $\underbrace{\text{sucesión decreciente}}_{\bigcap_{n_0=1}^{\infty} \bigcup_{n=n_0}^{\infty} \{ |x_n - x| > \varepsilon \}} \quad \forall \varepsilon > 0$

$$P\left(\bigcap_{n_0=1}^{\infty} \bigcup_{n=n_0}^{\infty} \{ |x_n - x| > \varepsilon \}\right) = \lim_{n_0 \rightarrow \infty} P\left(\bigcup_{n=n_0}^{\infty} \{ |x_n - x| > \varepsilon \}\right)$$

Nos falta b) \Rightarrow a)

$$P\left(\bigcap_{n_0=1}^{\infty} \bigcup_{n=n_0}^{\infty} |X_n - x| > \varepsilon\right) = 0 \quad \forall \varepsilon > 0.$$

Queremos ver que $P\left(\underbrace{X_n \rightarrow x}_{C^c}\right) = 0$.

$$P(C^c) = P\left(\bigcup_{M=1}^{\infty} \bigcap_{n_0=1}^{\infty} \bigcup_{n=n_0}^{\infty} \{|X_n - x| > 1/M\}\right)$$

(σ -subaditividad de P)

$$\leq \sum_{M=1}^{\infty} P\left(\bigcap_{n_0=1}^{\infty} \bigcup_{n=n_0}^{\infty} \{|X_n - x| > 1/M\}\right) = 0 \quad \therefore P(C^c) = 0 \checkmark$$

Criterio para asegurar convergencia c.t.p.

Lema 7.5

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. $X_n, X : \Omega \rightarrow \mathbb{R}$ variables aleatorias. Sea $A_n(\varepsilon) = \{|X_n - X| > \varepsilon\}$.

Si $\sum_{n=1}^{\infty} P(A_n(\varepsilon)) < \infty$ para todo $\varepsilon > 0$, entonces $X_n \xrightarrow{\text{c.t.p.}} X$.

dem: Por Borell - Cantelli(i), como la serie converge tenemos

$$\forall \varepsilon > 0, P\left(\limsup_n A_n(\varepsilon)\right) = 0 \Rightarrow \begin{matrix} \text{por Lema 7.4} \\ X_n \xrightarrow{\text{ctp}} X \end{matrix}$$

Definición 7.3 (Convergencia en Probabilidad)

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad. $X_n, X : \Omega \rightarrow \mathbb{R}$ variables aleatorias. Diremos que la sucesión de variables aleatorias $(X_n)_{n \geq 1}$ **converge en probabilidad** a la variable aleatoria X y escribimos $X_n \xrightarrow{P} X$ si para todo $\epsilon > 0$ tenemos que

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \epsilon) = 0 ,$$

Proposición 7.4

Tenemos entonces la siguiente implicación

$$X_n \xrightarrow{c.t.p.} X \Rightarrow X_n \xrightarrow{P} X$$

Demostración.

$$P(|X_{n_0} - X| > \varepsilon) \leq P(\cup_{n \geq n_0} |X_n - X| > \varepsilon) \rightarrow 0$$

cuando $n_0 \rightarrow \infty$ si $X_n \xrightarrow{c.t.p.} X$, por el Lema 7.4 c). □

La recíproca no vale. Veamos un contraejemplo.

Ejemplo 7.2

Sean $X_n \sim Be\left(\frac{1}{n}\right)$ independientes. Entonces $X_n \xrightarrow{P} 0$ pues

$$P(|X_n - 0| > \varepsilon) = P(|X_n| > \varepsilon) = P(X_n = 1) = \frac{1}{n} \rightarrow 0 \text{ cuando } n \rightarrow +\infty.$$

Pero $X_n \not\xrightarrow{c.t.p.} 0$. Calculemos $P(\limsup_{n \rightarrow \infty} \{|X_n| > \varepsilon\})$. Para eso, veamos

- $\sum_{n \geq 1} P(|X_n| > \varepsilon) = \sum_{n \geq 1} (1/n) = +\infty \forall \varepsilon > 0$.
- Además, los eventos $\{|X_n| > \varepsilon\}$ son independientes,

por el Lema de Borell - Cantelli (ii) tenemos que

$P(\limsup_{n \rightarrow \infty} \{|X_n| > \varepsilon\}) = 1$. Luego, por el Lema 7.4 resulta que $X_n \not\xrightarrow{c.t.p.} 0$.

¿Podría X_n converger c.t.p. a otro límite X ?

No, porque si $X_n \xrightarrow{c.t.p.} X$ entonces $X_n \xrightarrow{P} X$ y por lo tanto, $X = 0$ con probabilidad 1.

Convergencia en distribución, o débil

Podemos considerar una tercera forma de convergencia.

Definición 7.4

Sean X_n , X variables aleatorias. Decimos que X_n **converge en distribución** a la variable aleatoria X , y lo notaremos $X_n \xrightarrow{\mathcal{D}} X$, si para todo punto t de continuidad de F_X se tiene que

$$\lim_{n \rightarrow \infty} F_{X_n}(t) = F_X(t).$$

Es decir,

$$P(X_n \leq t) \xrightarrow{n \rightarrow \infty} P(X \leq t) \text{ para todo } t \text{ con } P(X = t) = 0.$$

¿Por qué pedir que la convergencia de las funciones de distribución sea válida sólo en los puntos de continuidad de la distribución límite? El siguiente ejemplo procura dar una respuesta a esta cuestión:

Ejemplo 7.3

Consideremos la sucesión de variables aleatorias dada por $W_n = 1/n$, es decir, son variables aleatorias constantes. Cualquiera sea la definición de convergencia es razonable que $W_n \rightarrow W = 0$. Notemos que $F_{W_n}(0) = 0$ para todo n mientras que $F_W(0) = 1$. Es decir, en el punto 0 no vale la convergencia de las acumuladas. Sin embargo, como F_W no es continua en el 0, la definición de convergencia débil excluye a este punto y, podemos garantizar que $W_n \xrightarrow{\mathcal{D}} W = 0$.

Observación 7.1

- ① La convergencia en distribución solo involucra a la función de distribución de las variables aleatorias. Observemos que *ni siquiera requiere que las variables estén definidas en un mismo espacio de probabilidad*, puesto que sólo considera el comportamiento de las respectivas funciones de distribución.
- ② La convergencia en distribución sirve principalmente para calcular probabilidades.

Para relacionar esta nueva noción de convergencia con las otras dos, tenemos el siguiente resultado:

Proposición 7.5

Vale la siguiente implicación:

$$X_n \xrightarrow{P} X \text{ entonces } X_n \xrightarrow{\mathcal{D}} X .$$

Para probar la implicación de la Proposición 7.5, nos será muy útil este lema auxiliar, que probaremos primero.

Lema 7.6 (Lema auxiliar)

Sean $X, Y : (\Omega, \mathcal{F}, P) \rightarrow \mathbb{R}$ variables aleatorias, sean $a > 0, \varepsilon > 0$.

Entonces

$$P(X \leq a) \leq P(Y \leq a + \varepsilon) + P(|X - Y| > \varepsilon).$$

$$\{X \leq a\} = \{X \leq a, Y \leq a + \varepsilon\} \cup \{X \leq a, Y > a + \varepsilon\}$$

$$\subset \{Y \leq a + \varepsilon\} \cup \{|X - Y| > \varepsilon\}$$

el resultado sigue de tomar probabilidades a esta relación y usar la aditividad

$$1) \{a + \varepsilon < Y(\omega)\} \Rightarrow X(\omega) < a < a + \varepsilon < Y(\omega)$$

$$2) \{X(\omega) < a\}$$



$$|X(\omega) - Y(\omega)| = Y(\omega) - X(\omega) > Y(\omega) - a > \varepsilon$$

Ahora probamos la proposición

Sea t un punto de continuidad de F_X . Queremos ver que

$F_{X_n}(t) \xrightarrow{n \rightarrow \infty} F_X(t)$. Sea $\varepsilon > 0$. Usamos el Lema aux:

$$1) F_{X_n}(t) \leq F_X(t+\varepsilon) + P(|X_n - X| > \varepsilon)$$

$x = X_n$
 $y = X$
 $a = t$

$\stackrel{\downarrow}{}$

$$\stackrel{\text{"}}{P(X_n \leq t)} \quad P(X \leq t+\varepsilon)$$

Lema aux

$$2) F_X(t-\varepsilon) \leq F_{X_n}(t) + P(|X_n - X| > \varepsilon)$$

$x = X$
 $y = X_n$
 $a = t-\varepsilon$

$\stackrel{\text{"}}{P(X \leq t-\varepsilon)} \quad P(X_n \leq t)$

Entonces:

$$F_X(t-\varepsilon) - P(|X_n - X| > \varepsilon) \leq F_{X_n}(t) \leq F_X(t+\varepsilon) + P(|X_n - X| > \varepsilon)$$

$\uparrow \text{por 2}) \quad \uparrow \text{por 1})$

$X_n \xrightarrow{P} X$ por lo que, sabemos que $P(|X_n - X| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \varepsilon > 0$

Copiamos: por 2) por 1)
↓ ↓

$$F_X(t - \varepsilon) - P(|X_n - X| > \varepsilon) \leq F_{X_n}(t) \leq F_X(t + \varepsilon) + P(|X_n - X| > \varepsilon)$$

Tomando $\liminf_{n \rightarrow \infty}$ y $\limsup_{n \rightarrow \infty}$ tenemos:

$$F_X(t - \varepsilon) \leq \liminf_{n \rightarrow \infty} F_{X_n}(t) \leq \limsup_{n \rightarrow \infty} F_{X_n}(t) \leq F_X(t + \varepsilon)$$

Pero F_X es continua en t , tomando $\lim_{\varepsilon \rightarrow 0}$ tenemos

$$F_X(t) = \lim_{n \rightarrow \infty} F_{X_n}(t) \quad \checkmark$$

Mostremos un ejemplo de que la recíproca es falsa.

Ejemplo 7.4

Sean $X \sim \text{Be}(1/2)$. Sean $X_n = X \forall n \in \mathbb{N}$. Claramente las X_n no son independientes. Sea $Y = 1 - X$, entonces $Y \sim \text{Be}(1/2)$. Entonces

$X_n \xrightarrow{\mathcal{D}} Y$, pues $F_{X_n}(t) = F_Y(t) \forall t \in \mathbb{R}, \forall n \in \mathbb{N}$.

Sin embargo, $X_n \not\xrightarrow{P} Y$ pues:

$$|X_n - Y| = 1 \quad \forall n \in \mathbb{N}.$$

Luego $P(|X_n - Y| > \varepsilon) = 1 \forall \varepsilon < 1, \forall n \in \mathbb{N}$.

Espacios \mathcal{L}^p

Definición 7.5 (Espaces \mathcal{L}^p)

Dado un espacio de probabilidad (Ω, \mathcal{F}, P) , para $p \geq 1$, $\mathcal{L}^p = \mathcal{L}^p(\Omega, \mathcal{F}, P)$ denota el conjunto de todas las variables aleatorias definidas en Ω con $E(|X|^p) < \infty$.

\mathcal{L}^1 es el conjunto de variables aleatorias con esperanza finita. A \mathcal{L}^2 lo definimos antes, cuando definimos varianza. Si p es entero, a la $E(X^p)$ se la denomina p -ésimo momento de la variable aleatoria X .

Definición 7.6 (convergencia en \mathcal{L}^p)

Sean $(X_n)_{n \geq 1}$, X variables aleatorias definidas en el mismo espacio de probabilidad (Ω, \mathcal{F}, P) , $p \geq 1$, $X_n, X \in \mathcal{L}^p$. Diremos que $(X_n)_{n \geq 1}$ converge a X en \mathcal{L}^p , y lo notaremos $X_n \xrightarrow{\mathcal{L}^p} X$ si

$$E(|X_n - X|^p) \rightarrow 0 \text{ cuando } n \rightarrow +\infty$$

Lema 7.7

Si $1 \leq p < q$, entonces $\mathcal{L}^q \subset \mathcal{L}^p$.

Consecuencia probabilística: Si existe el momento q de una variable aleatoria, $E(|X|^q) < +\infty$ entonces existen todos los momentos menores.

(sin demo) Se usa usando la desigualdad de Hölder

Si la asumimos, o sea si tenemos x con $E|x|^p < \infty$ y y con $E|y|^q < \infty$ entonces $1 \leq p < q \Rightarrow 1 < q/p$. Luego $g(t) = |t|^{q/p}$ es convexa

La desigualdad de Jensen nos da: $g(E(w)) \leq E(g(w))$
si w s.t. $E(w)$ y $E(g(w))$ sean finitas.

Si la aplicamos a $w = |x|^p$ tenemos

$$(E|x|^p)^{q/p} \leq E(|x|^q)$$

Proposición 7.6

Sean $(X_n)_{n \geq 1}$, X variables aleatorias definidas en (Ω, \mathcal{F}, P) . Vale la siguiente implicación

$$X_n \xrightarrow{L^p} X \text{ entonces } X_n \xrightarrow{P} X$$

dem: Sea $\epsilon > 0$.

$$P(|X_n - X| > \epsilon) \leq \frac{E(|X_n - X|^p)}{\epsilon^p} \xrightarrow[n \rightarrow \infty]{\quad} 0 \quad \text{y listo}$$

↑
desigualdad de Markov

con $\varphi(x) = x^p$ es creciente si $p \geq 1$

Relación entre los distintos tipos convergencias

$$\begin{array}{ccc} x_n \xrightarrow{\text{c.s.}} x & \Rightarrow & x_n \xrightarrow{P} x \rightarrow x_n \xrightarrow{\otimes} x \\ x_n \xrightarrow{L^P} x & \Rightarrow & \end{array}$$

- En algunos casos particulares (agregando hipótesis) salen algunas reciprocas parciales
- Otras preguntas de interés
 - ¿alguna implica la conv de las esperanzas?
 - $g(x_n) \rightarrow g(x)$ en algún sentido?
 - $x_n + y_n \rightarrow x + y$ idem?

Ejemplo 7.5 (otro de convergencia en proba pero no c.s.)

En el espacio de probabilidad $([0, 1], \mathcal{B}, |\cdot|)$, (la probabilidad está dada por la longitud) definamos las variables aleatorias X_1, X_2, \dots por

$$\begin{aligned}X_1 &\equiv 1 \\X_2 &= I_{[0,1/2]} \\X_3 &= I_{(1/2,1]} \\X_4 &= I_{[0,1/4]} \\X_5 &= I_{(1/4,1/2]} \cdots \\X_8 &= I_{[0,1/8]} \cdots\end{aligned}$$

Observemos que X_n converge a cero en probabilidad, puesto que $0 < \varepsilon < 1$ $P\{|X_n - 0| \geq \varepsilon\} = P\{X_n \neq 0\} \rightarrow 0$. Por otro lado, no hay ningún x para el cual $X_n(x) \rightarrow 0$, puesto que para todo x , hay infinitos n para los cuales $X_n(x) = 1$, e infinitos n para los cuales $X_n(x) = 0$. De hecho, para cada x y cada k hay exactamente un n entre 2^k y $2^{k+1} - 1$ para el cual $X_n(x) = 1$. Por lo tanto, la sucesión $X_n(x)$ no converge para ningún x , y por lo tanto X_n no converge c.s.

Algunas recíprocas parciales

Proposición 7.7 (teo de la convergencia acotada)

Si $X_n \xrightarrow{P} X$ y existe C tal que $|X_n| \leq C$, $|X| \leq C$. Entonces, $X_n \xrightarrow{L^1} X$.

dem : sea $\epsilon > 0$

$$|X_n - X| \leq \epsilon I_{\{|X_n - X| \leq \epsilon\}} + 2C I_{\{|X_n - X| > \epsilon\}}$$

$$\begin{aligned} \Rightarrow E|X_n - X| &\leq \epsilon P(|X_n - X| \leq \epsilon) + 2C P(|X_n - X| > \epsilon) \\ &\leq \epsilon + 2C \underbrace{P(|X_n - X| > \epsilon)}_{\rightarrow 0 \text{ } n \rightarrow \infty} \end{aligned}$$

$\Rightarrow \limsup_{n \rightarrow \infty} E|X_n - X| \leq \epsilon \quad \forall \epsilon > 0$ y el resultado se obtiene tomando límite cuando $\epsilon \rightarrow 0$.

La Proposición 7.7 deja de ser cierta si no disponemos de algún tipo de cota para las variables, como muestra el siguiente ejercicio.

Ejercicio 7.1

Consideremos $Y_n = nI_{[0,1/n]}(U_n)$, para U_n variables uniformes en el intervalo $[0, 1]$. ¿Qué puede decir sobre el límite en probabilidad de Y_n ? ¿Qué puede decir sobre la sucesión $E[Y_n]$?

Lema 7.8

Si $X_n \xrightarrow{L^p} X$ y $1 \leq \ell < p$, entonces $X_n \xrightarrow{L^\ell} X$. Además, $E[X_n] \rightarrow E[X]$.

Demostración.

Siendo $\ell < p$, $g(x) = |x|^{p/\ell}$ es una función convexa. De la desigualdad de Jensen, $(g(E[W])) \leq E[g(W)]$) tenemos que

$$\left(E [|X_n - X|^\ell] \right)^{p/\ell} \leq E [|X_n - X|^{\ell p / \ell}] \leq E [|X_n - X|^p]$$

Luego, si $(X_n)_{n \geq 1}$ converge en L^p , tenemos que vale la convergencia en L^1 y por consiguiente,

$$|E[X_n] - E[X]| \leq E [|X_n - X|], \quad (\text{por el Lema 6.6})$$

de donde concluimos que $E[X_n] \rightarrow E[X]$. □

Teorema 7.9

Si $X_n \xrightarrow{P} X$ en probabilidad, entonces existe una subsucesión $X_{n_k} \xrightarrow{c.s.} X$.

Demostración.

Primero construimos la subsucesión n_k . Luego vamos a usar el criterio que garantiza la convergencia casi segura dado en el Lema 7.5: si

$\sum_{k=1}^{\infty} P(\{|X_{n_k} - X| > \varepsilon\}) < \infty$ para todo $\varepsilon > 0$, entonces $X_{n_k} \xrightarrow{c.s.} X$.

Basta tomar n_k de forma tal que

$$P\left(|X_{n_k} - X| \geq \frac{1}{k}\right) < \frac{1}{2^k}.$$

En tal caso, dado $\varepsilon > 0$, tomemos k_0 de forma tal que $1/k < \varepsilon$, para $k \geq k_0$. Tenemos entonces que

$$\{|X_{n_k} - X| \geq \varepsilon\} \subset \{|X_{n_k} - X| \geq 1/k\} \text{ para todo } k \geq k_0.$$

Luego $P(|X_{n_k} - X| \geq \varepsilon) \leq P(|X_{n_k} - X| \geq 1/k)$ para todo $k \geq k_0$. □

Demostración.

Con lo cual

$$\begin{aligned} \sum_{k=1}^{\infty} P(|X_{n_k} - X| \geq \varepsilon) &\leq (k_0 - 1) + \sum_{k=k_0}^{\infty} P(|X_{n_k} - X| \geq 1/k) \\ &< (k_0 - 1) + \sum_{k \geq k_0} \frac{1}{2^k} < \infty \end{aligned}$$



El siguiente resultado es un ejercicio de la práctica

Teorema 7.10 (sub de la sub)

Una sucesión de variables aleatorias $(X_n)_{n \in \mathbb{N}}$ converge en probabilidad a una variable aleatoria X si y sólo si toda subsucesión de $(X_n)_{n \in \mathbb{N}}$ contiene otra subsucesión que converge casi seguramente a X

Recordemos este resultado

Lema 7.11 (de espacios métricos)

Sea $(y_n)_{n \in \mathbb{N}}$ una sucesión de elementos en un espacio vectorial. Si toda subsucesión $(y_{n_j})_{j \in \mathbb{N}}$ tiene una subsucesión $(y_{n_j})_{j \in \mathbb{N}}$ que converge a y , entonces $y_n \rightarrow y$.

La prueba del Teorema 7.10 sale usando este Lema de espacios métricos para la sucesión $y_n = P(|X_n - X| > \varepsilon)$. Recordar también el Teorema 7.9.

Probaremos el lema de espacios métricos.

→ (ejercicio de la práctica)

Si $y_n \rightarrow y \Rightarrow \exists G$ abierto que contiene a y y una subsucesión y_{n_j} con $y_{n_j} \notin G \forall j$. Pero entonces ninguna subsucesión de y_{n_j} convergerá a y , lo cual es absurdo.

Obs: La caracterización de la convergencia en probabilidad que se obtiene a través del Teorema 7.10 es muy útil para probar resultados (e.g., ciertas operaciones mantienen la convergencia en probab.).

Vimos que la convergencia en distribución es la forma más débil de convergencia pues sólo involucra las funciones de distribución y no hace referencia al espacio de probabilidades subyacente. Sin embargo, el próximo resultado nos da una construcción muy útil para vincularla con la convergencia casi segura.

Teorema 7.12 (Teorema de representación de Skorokhod)

Si $X_n \xrightarrow{\mathcal{D}} X$, entonces existe (Ω, \mathcal{F}, P) espacio de probabilidad, $(Y_n)_{n \geq 1}$, Y , variables aleatorias definidas en (Ω, \mathcal{F}, P) , de forma tal que

$$X_n \sim Y_n, \quad X \sim Y \quad \text{y} \quad Y_n \xrightarrow{c.s.} Y.$$

Para demostrar el Teorema precedente, haremos uso del siguiente resultado:

Lema 7.13

Sean $(F_n)_{n \geq 1}$, F funciones de distribución tales que

$$F_n(u) \xrightarrow{n \rightarrow \infty} F(u) \text{ para todo } u \text{ punto de continuidad de } F.$$

Entonces, $F_n^{-1}(u) \rightarrow F^{-1}(u) \forall u \in (0, 1)$, punto de continuidad de F^{-1} .

Demostración.

Sea $u \in (0, 1)$, punto de continuidad de F^{-1} . Veremos que

$$1) \liminf F_n^{-1}(u) \geq F^{-1}(u) \quad y \quad 2) \limsup F_n^{-1}(u) \leq F^{-1}(u)$$

Recordemos la definición de F^{-1}

$$F^{-1}(u) = \inf A_u, \text{ con } A_u = \{x \in \mathbb{R} : F(x) \geq u\}.$$

Vimos que $x < F^{-1}(u) \Leftrightarrow F(x) < u$. (Lema 4.10.), $F^{-1}(u) \leq x \Leftrightarrow u \leq F(x)$ □

Demostración.

Tomemos $\varepsilon > 0$ y x punto de continuidad de F en el intervalo $(F^{-1}(u) - \varepsilon, F^{-1}(u))$. En tal caso, tenemos que $x < F^{-1}(u)$ y por consiguiente, $F(x) < u$.

Como $F_n(x) \rightarrow F(x)$, existe n_0 a partir del cual $F_n(x) < u$, para todo $n \geq n_0$.

Luego, tenemos que $x < F_n^{-1}(u)$ para $n \geq n_0$, y por consiguiente

$$F^{-1}(u) - \varepsilon < x \leq \liminf_{n \rightarrow \infty} F_n^{-1}(u) \quad \forall \varepsilon > 0,$$

de donde concluimos que

$$F^{-1}(u) \leq \liminf_{n \rightarrow \infty} F_n^{-1}(u).$$

(Para demostrar 1), no usamos que F^{-1} es continua en u). □

Seguimos con la demostración.

Para demostrar 2), fijemos $u' > u$ y $\epsilon > 0$. Tomemos y punto de continuidad de F , con $F^{-1}(u') < y < F^{-1}(u') + \epsilon$.

Luego, tenemos que $u < u' \leq F(y)$. (por Lema 4.10)

Como $F_n(y) \rightarrow F(y)$, para n suficientemente grande, tenemos que $u < F_n(y)$ y por consiguiente, $F_n^{-1}(u) \leq y < F^{-1}(u') + \epsilon$. Tenemos entonces que

$$\limsup_{n \rightarrow \infty} F_n^{-1}(u) \leq F^{-1}(u') + \epsilon, \forall \epsilon > 0, \forall u' > u.$$

Tomando $\epsilon \rightarrow 0$, obtenemos que

$$\limsup_{n \rightarrow \infty} F_n^{-1}(u) \leq F^{-1}(u'), \forall u' > u$$

y usando ahora la continuidad de F^{-1} en u , haciendo $u' \downarrow u$, deducimos que

$$\limsup_{n \rightarrow \infty} F_n^{-1}(u) \leq F^{-1}(u). \quad \checkmark \quad 2)$$

Corolario 7.14

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad, $U : \Omega \rightarrow \mathbb{R}$ una variable aleatoria con distribución $U \sim \mathcal{U}(0, 1)$. Entonces,

$$F_n^{-1}(U) \xrightarrow{c.s.} F^{-1}(U).$$

Demostración.

Denotemos con $C = \{u \in (0, 1) : F^{-1} \text{ no es continua en } u\}$. Como F^{-1} es creciente (en sentido amplio), C es a lo sumo numerable. Luego, $P(U \in C) = 0$. Tenemos entonces que $\Omega \setminus U^{-1}(C)$ es un conjunto de probabilidad uno, donde $F_n^{-1}(U) \rightarrow F^{-1}(U)$.

□

Volvamos al Teorema 7.12.

Teorema 7.12 (Teorema de representación de Skorokhod)

Si $X_n \xrightarrow{\mathcal{D}} X$, entonces existe (Ω, \mathcal{F}, P) espacio de probabilidad, $(Y_n)_{n \geq 1}$, Y , variables aleatorias definidas en (Ω, \mathcal{F}, P) , de forma tal que

$$X_n \sim Y_n, \quad X \sim Y \quad \text{y} \quad Y_n \xrightarrow{\text{c.s.}} Y.$$

Demostración.

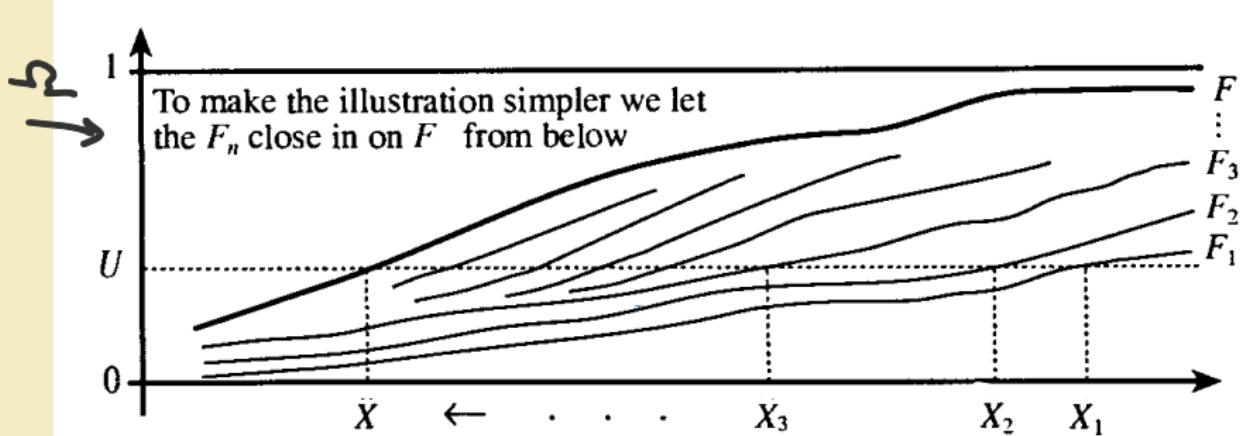
Si ponemos $Y_n = F_n^{-1}(U)$, $Y = F^{-1}(U)$, tenemos que $\underline{Y_n} \sim F_n$, $\underline{Y} \sim F$ y además

$$F = F_X \quad F_n = F_{X_n} \quad Y_n \xrightarrow{\text{c.s.}} Y.$$



La figura (del libro Coupling, Stationarity and Regeneration, Thorisson) ilustra la construcción del Teorema. Para simplificar la ilustración, el autor grafica las funciones $F_n \leq F_{n+1} \leq \dots F$

En el eje y elige un valor de U , y en el eje x representa la sucesión $F_n^{-1}(U) = X_n$ que, al ser las F_n crecientes, converge a $x = F^{-1}(U)$ de forma decreciente



Este es un ejemplo de un acoplamiento (coupling).

Convergencia en distribución

Lema 7.15

Sean X Y variables aleatorias con misma distribución: $X \sim Y$. Entonces, $g(X) \sim g(Y)$ para toda función g medible.

Demostración.

$$P(g(X) \in A) = P(X \in g^{-1}(A)) = P(Y \in g^{-1}(A)) = P(g(Y) \in A)$$



Lema 7.16

$X_n \xrightarrow{\mathcal{D}} X$ y $P(X = c) = 1$ para $c \in \mathbb{R}$, entonces $X_n \xrightarrow{P} c$.

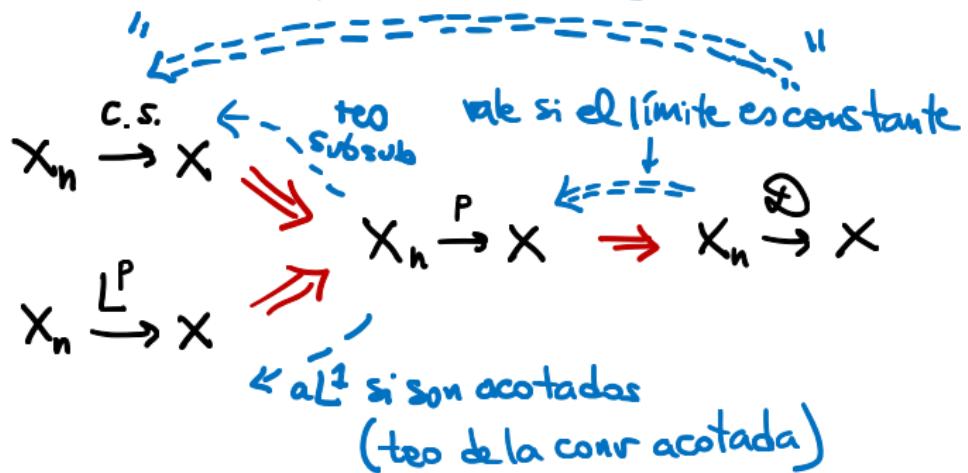
Demostración.

Fijemos $\epsilon > 0$. $F_X(t)$ es continua para todo $t \neq c$. Entonces

$$\begin{aligned} P(|X_n - c| > \epsilon) &= P(X_n < c - \epsilon) + P(X_n > c + \epsilon) \\ &\leq P(X_n \leq c - \epsilon) + P(X_n > c + \epsilon) \\ &= F_{X_n}(c - \epsilon) + 1 - F_{X_n}(c + \epsilon) \\ &\rightarrow F_X(c - \epsilon) + 1 - F_X(c + \epsilon) \\ &= 0 + 1 - 1 = 0 \end{aligned}$$



Represent. Skorokhod



Relaciones entre las convergencias (y sus reciprocas parciales)

Además si $x_n \xrightarrow{P} x \Rightarrow x_n \xrightarrow{L^1} x \Rightarrow E(x_n) \rightarrow E(x)$.

Nos falta un resultado que vincula corr en distrib y esperanzas

Convergencia de vectores aleatorios

Definición 7.7

Sean $\tilde{\mathbf{X}}_n$, $\tilde{\mathbf{X}} \in \mathbb{R}^k$ vectores aleatorios, $n \in \mathbb{N}$, $\tilde{\mathbf{X}}_n = (X_n(1), \dots, X_n(k))$ y $\tilde{\mathbf{X}} = (X(1), \dots, X(k))$.

- a) Decimos que $\tilde{\mathbf{X}}_n \xrightarrow{P} \tilde{\mathbf{X}}$ si para todo $\varepsilon > 0$ se tiene que $P(\|\tilde{\mathbf{X}}_n - \tilde{\mathbf{X}}\| > \varepsilon) \rightarrow 0$ cuando $n \rightarrow \infty$.
- b) Decimos que $\tilde{\mathbf{X}}_n \xrightarrow{c.s.} \tilde{\mathbf{X}}$ cuando $P\left(\left\{\omega \in \Omega : \tilde{\mathbf{X}}_n(\omega) \rightarrow \tilde{\mathbf{X}}(\omega)\right\}\right) = 1$.

Afortunadamente el siguiente resultado facilita el tratamiento de estas convergencias.

Proposición 7.8

- i. $\tilde{X}_n \xrightarrow{P} \tilde{X} \iff X_n(i) \xrightarrow{P} X(i) \quad \forall i = 1, \dots, k.$
- ii. $\tilde{X}_n \xrightarrow{c.s.} \tilde{X} \iff X_n(i) \xrightarrow{c.s.} X(i) \quad \forall i = 1, \dots, k.$

(i) ejercicio. (Sug: $|X_n(i) - X(i)| \leq |\tilde{X}_n - \tilde{X}| \leq \sum_{j=1}^k |X_n(j) - X(j)|$)

Además, probar (usando complementos) y usar que

$$\{|x| + |y| > \varepsilon\} \subset \{|x| > \varepsilon/2\} \cup \{|y| > \varepsilon/2\}$$

(ii) Sean $B(i) = \{\omega \in \Omega : X_n(i)(\omega) \rightarrow X(i)(\omega)\}$ $A = \{\tilde{X}_n \rightarrow \tilde{X}\}$

Entonces $A \subset B(i) \quad \forall i = 1, \dots, k$ por (A). y $\bigcap_{i=1}^k B(i) \subset A$ por (B)

\Rightarrow Tenemos $P(A) = 1 \Rightarrow P(B(i)) = 1 \quad \forall i = 1, \dots, k \quad \checkmark$

\Leftarrow Sabemos que $P(B(i)) = 1, 1 \leq i \leq k$, y la $P(\bigcap_{i=1}^k B(i)) = 1 \leq P(A) \quad \checkmark$
 (intersección finita de conjuntos de probab 1 tiene probab 1)

Convergencias y funciones continuas

Las convergencias casi segura, en probabilidad y distribución se preservan bajo funciones continuas.

Teorema 7.17

Sean $g : \mathbb{R}^k \rightarrow \mathbb{R}$ continua, $h : \mathbb{R} \rightarrow \mathbb{R}$ continua. $(\tilde{\mathbf{X}}_n)_{n \geq 1}$, $\tilde{\mathbf{X}}$ vectores aleatorios, $(\mathbf{X}_n)_{n \geq 1}$, \mathbf{X} variables aleatorias.

① $\tilde{\mathbf{X}}_n \xrightarrow{c.s.} \tilde{\mathbf{X}}$ entonces $g(\tilde{\mathbf{X}}_n) \xrightarrow{c.s.} g(\tilde{\mathbf{X}})$.

② $\tilde{\mathbf{X}}_n \xrightarrow{P} \tilde{\mathbf{X}}$ entonces $g(\tilde{\mathbf{X}}_n) \xrightarrow{P} g(\tilde{\mathbf{X}})$.

③ $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$ entonces $h(\mathbf{X}_n) \xrightarrow{D} h(\mathbf{X})$. \mathbf{X}_n, \mathbf{X} variables aleat

dem 1) $A = \{w \in \Omega : \tilde{\mathbf{X}}_n(w) \rightarrow \tilde{\mathbf{X}}(w)\} \subseteq \{w \in \Omega : g(\tilde{\mathbf{X}}_n) \rightarrow g(\tilde{\mathbf{X}})\}$

como $P(A) = 1$, tenemos probado el resultado.

Sabemos

2) $x_n \xrightarrow{P} x$. Queremos ver que $g(x_n) \xrightarrow{P} g(x) \Leftrightarrow$

(usando el teo de la sub sub) de cada subsucesión $g(x_{n_j})$, puede extraerse una subsubsucesión $g(x_{n_j,h})$ que converge casi seguramente.

Sea $g(x_{n_j})$ una subsucesión arbitraria. Sabemos que $x_n \xrightarrow{P} x$.
Luego, sabemos, por el teo de la sub de la sub, que existe una subsucesión de (x_{n_j}) , digámosla $(x_{n_j,h})_h$ que converge c.s. a $x \Rightarrow$ por 1) $g(x_{n_j,h}) \xrightarrow{\text{c.s.}} g(x)$.

Ahora, usando nuevamente el teo de la sub de la sub, obtenemos que $g(x_n) \xrightarrow{P} g(x) \checkmark$

3) $X_n \xrightarrow{\mathcal{D}} X$. Usando el teo de Representación de Skorokhod, sabemos que existen un espacio (Ω, \mathcal{F}, P)

y va: $Y_n \sim X_n$, $Y \sim X$ tq: $Y_n \xrightarrow{c.s.} Y$

$$\Rightarrow h(Y_n) \xrightarrow[\text{por } \uparrow]{c.s.} h(Y) \Rightarrow h(Y_n) \xrightarrow{\mathcal{D}} h(Y)$$

por jerarquía de conv.

Luego, como $h(Y) \sim h(X)$ (Lema 7.15) ✓

En particular, tenemos que

Corolario 7.18

Si $X_n \xrightarrow{\text{c.s.}} X$ e $Y_n \xrightarrow{\text{c.s.}} Y$, entonces



1) $X_n + Y_n \xrightarrow{\text{c.s.}} X + Y$

2) $X_n \cdot Y_n \xrightarrow{\text{c.s.}} X \cdot Y$

Y también si $X_n \xrightarrow{P} X$ e $Y_n \xrightarrow{P} Y$, entonces

3) $X_n + Y_n \xrightarrow{P} X + Y$

4) $X_n \cdot Y_n \xrightarrow{P} X \cdot Y$

Finalmente, si $X_n \xrightarrow{\mathcal{D}} X$, entonces

5) $aX_n \xrightarrow{\mathcal{D}} aX$, y

6) $X_n + a \xrightarrow{\mathcal{D}} X + a$.

Convergencia en distribución

En general no vale que si $X_n \xrightarrow{\mathcal{D}} X$ e $Y_n \xrightarrow{\mathcal{D}} Y$, entonces $X_n + Y_n \xrightarrow{\mathcal{D}} X + Y$. ¿Qué podemos decir de la convergencia en distribución de la suma o del producto de dos sucesiones de variables aleatorias que convergen en distribución?

Para responder a esto necesitamos definir y probar algunas cosas antes.

Definición 7.8 (acotada en probabilidad)

Una colección de variables aleatorias $\{X_\alpha : \alpha \in J\}$ se dice **acotada en probabilidad** o “tight” si para todo $\varepsilon > 0$ existe $M > 0$ de forma tal que

$$P(|X_\alpha| > M) < \varepsilon \quad \forall \alpha \in J.$$

Por ejemplo, $(X_n)_{n \geq 1}$ dada por $X_n = n$ no es tight.

Observación 7.2

Toda variable aleatoria es acotada en probabilidad (ejercicio).

Ejercicio 7.2

Para las siguientes familias, dar condiciones necesarias y suficientes para que $(X_n)_{n \geq 1}$ sea acotado en probabilidad.

- ① $X_n \sim \mathcal{N}(\mu_n, 1)$.
- ② $X_n \sim \mathcal{N}(0, \sigma_n^2)$.
- ③ $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$.
- ④ $X_n \sim \mathcal{U}(0, a_n)$.

Lema 7.19

Si $X_n \xrightarrow{P} X$, entonces $(X_n)_{n \geq 1}$ es acotada en probabilidad.

La prueba es un ejercicio de la práctica.

Lema 7.20

Si $X_n \xrightarrow{\mathcal{D}} X$, entonces $(X_n)_{n \geq 1}$ es acotada en probabilidad.

La prueba queda como ejercicio.

Lema 7.21

Si $X_n \xrightarrow{P} 0$ e $(Y_n)_{n \geq 1}$ es acotada en probabilidad, entonces $X_n Y_n \xrightarrow{P} 0$.

Sea $\delta > 0$. Existe $M > 0$: $P(|Y_n| > M) < \delta$ $\forall n \geq n_0$. Sea $\varepsilon > 0$

$$\{|X_n Y_n| > \varepsilon\} = \{|X_n Y_n| > \varepsilon \cap |Y_n| > M\} \cup \underbrace{\quad}_{\text{disjunta}}$$

$$\{|X_n Y_n| > \varepsilon \cap |Y_n| \leq M\}$$

$$\subseteq \{|Y_n| > M\} \cup \{|X_n| > \varepsilon/M\}$$

Luego, $0 \leq P(|X_n Y_n| > \varepsilon) \leq \underbrace{P(|Y_n| > M)}_{< \delta} + P(|X_n| > \varepsilon/M)$

$$< \delta + P(|X_n| > \varepsilon/M)$$

tenemos $\limsup_{n \rightarrow \infty}$

$0 \leq \limsup_{n \rightarrow \infty} P(|X_n Y_n| > \varepsilon) \leq \delta + \delta > 0$ y el resultado surge de hacer tender δ a 0.

Teorema 7.22 (Teorema de Slutsky)

Sean $X_n \xrightarrow{\mathcal{D}} X$ e $Y_n \xrightarrow{P} c \in \mathbb{R}$. Entonces

a) $X_n + Y_n \xrightarrow{\mathcal{D}} X + c$

b) $X_n \cdot Y_n \xrightarrow{\mathcal{D}} X \cdot c$

dem: Asumamos que lo hemos probado para $c=0$, y veamos cómo deducir el resultado cuando $c \neq 0$.

a) Sabemos que $Y_n - c \xrightarrow{\mathcal{D}} 0$ por Corolario 7.18 G)

Luego, por el caso $c=0$, tendremos $X_n + (Y_n - c) \xrightarrow{\mathcal{D}} X$

Finalmente $X_n + (Y_n - c) + c = X_n + Y_n \xrightarrow{\mathcal{D}} X + c$ por Coro 7.18 G)

b) Por el caso $c=0$, resulta $X_n(Y_n - c) \xrightarrow{\mathcal{D}} 0$.

Es decir $X_n Y_n - c X_n = W_n \xrightarrow{\mathcal{D}} 0$. Como el límite es constante, resulta que $W_n \xrightarrow{P} 0$. Pero como $Z_n = c X_n \xrightarrow{\mathcal{D}} c X$, listo usando a) para coro 7.18 G) W_n y Z_n

Demostración.

a) Probemos el caso $c = 0$, si $X_n \xrightarrow{\mathcal{D}} X$ e $Y_n \xrightarrow{P} 0$, entonces $X_n + Y_n \xrightarrow{\mathcal{D}} X$.

Sea x punto de continuidad de F_X .

$$\begin{aligned} P(X_n + Y_n \leq x) &\leq P(X_n + Y_n \leq x, |Y_n| > \varepsilon) + P(X_n + Y_n \leq x, |Y_n| \leq \varepsilon) \\ &\leq P(|Y_n| > \varepsilon) + P(X_n \leq x + \varepsilon) \end{aligned}$$

Tomando ε tal que $x + \varepsilon$ sea un punto de continuidad de F_X obtenemos que

$$\limsup_{n \rightarrow \infty} F_{X_n + Y_n}(x) \leq F_X(x + \varepsilon),$$

y usando la continuidad a derecha de F_X , concluimos que

$$\limsup_{n \rightarrow \infty} F_{X_n + Y_n}(x) \leq F_X(x)$$



Demostración.

Por otra parte, tenemos que

$$P(X_n + Y_n \leq x) \geq P(X_n \leq x - \varepsilon \cap |Y_n| \leq \varepsilon),$$

Luego, eligiendo ε tal que $x - \varepsilon$ sea punto de continuidad de F_X conluimos que

$$\liminf_{n \rightarrow \infty} F_{X_n + Y_n}(x) \geq F_X(x - \varepsilon).$$

El resultado se deduce haciendo ε ir a cero, y usando que x es punto de continuidad de F_X .



b) Falta probar la convergencia del producto en el caso $c=0$.

$X_n \xrightarrow{d} X$. Como la convergencia en distribución garantiza que la sucesión $(X_n)_{n \geq 1}$ sea acotada en probabilidad, tenemos

$$Y_n \xrightarrow{P} 0 \text{ y } (X_n)_{n \geq 1} \text{ acot en probab} \Rightarrow X_n Y_n \xrightarrow{P} 0 \Rightarrow X_n Y_n \xrightarrow{d} 0$$

Lema 7.21



Teorema 7.23 (caracterización de la conv en distrib vía esperanzas)

Son equivalentes:

a) $X_n \xrightarrow{\mathcal{D}} X$

b) $E[g(X_n)] \rightarrow E[g(X)]$ para toda g función continua y acotada.

a) \Rightarrow b) Sabemos $X_n \xrightarrow{\mathcal{D}} X$. Sea g cont y acotada

Por el Teorema de Skorokhod, $\exists (\Omega, \mathcal{F}, P)$ espacio de probabilidad y variables Y_n, Y definidas en él tales que $Y_n \sim X_n$, $Y \sim X$ e $Y_n \xrightarrow{c.s.} Y \Rightarrow g(Y_n) \xrightarrow{c.s.} g(Y) \Rightarrow g(Y_n) \xrightarrow{P} g(Y)$ (jerarquía de conv)

Como g es acotada, x el teo de la conv acotada, resulta:

$g(Y_n) \xrightarrow{L^1} g(Y) \Rightarrow E(g(Y_n)) \rightarrow E(g(Y))$. Como (Lema 7.15)
por Lema 7.8

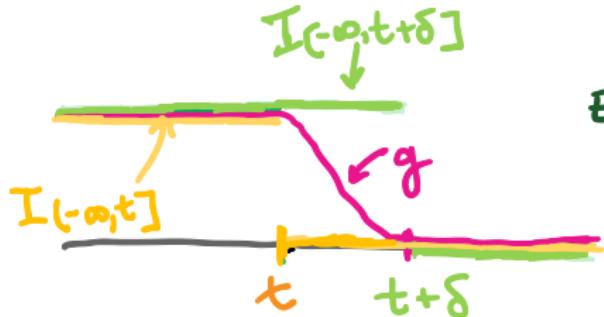
$E(g(Y_n)) = E(g(X_n))$ y $E(g(Y)) = E(g(X))$ queda probado ✓

b) \Rightarrow a) Sabemos que $\forall g: \mathbb{R} \rightarrow \mathbb{R}$ continua y acotada vale que $E(g(x_n)) \rightarrow E(g(x))$.

Queremos ver que $x_n \xrightarrow{\text{D}} x$, ie $F_{x_n}(t) \rightarrow F_x(t)$ $\forall t$ punto de continuidad de F_x .
 Pero: $F_x(t) = E(I_{(-\infty, t]}(x))$ así que la cuenta

consiste en elegir una función g acotada y continua, que se "parezca" a la $I_{(-\infty, t]}$. Sea t un punto de continuidad de F_x

Dado $\delta > 0$, sea $g: \mathbb{R} \rightarrow \mathbb{R}$ continua y acotada tal que:



$$I_{(-\infty, t]}(x) \leq g(x) \leq I_{(-\infty, t+\delta]}(x)$$

Entonces

$$I_{(-\infty, t]}(x_n) \leq g(x_n) \quad \text{y también:}$$

$$g(x) \leq I_{(-\infty, t+\delta]}(x)$$

Copiamos

$$I_{(-\infty, t]}(x_n) \leq g(x_n) \quad y \quad g(x) \leq I_{(-\infty, t+\delta]}(x)$$

Tomando esperanza, tenemos

$$F_{X_n}(t) = P(X_n \leq t) \leq E(g(x_n)) \quad y \quad E(g(x)) \leq F_X(t+\delta)$$

Tomamos $\limsup_{n \rightarrow \infty}$ y obtenemos:

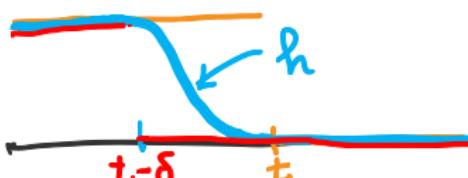
usamos la hip.

$$\limsup_{n \rightarrow \infty} F_{X_n}(t) \leq \lim_{n \rightarrow \infty} E(g(x_n)) = E(g(x)) \leq F_X(t+\delta) \quad (4)$$

Ahora dejamos que $\delta > 0$, obtenemos: (\times continuidad a derecha de F_X)

$$\limsup_{n \rightarrow \infty} F_{X_n}(t) \leq F_X(t)$$

Sea ahora h continua y acotada:



$$\begin{aligned} I_{(-\infty, t-\delta]}(x) &\leq h(x) \leq I_{(-\infty, t]}(x) \\ \rightarrow I_{(-\infty, t-\delta]}(x) &\leq h(x) \quad \text{y} \quad h(x_n) \leq I_{(-\infty, t]}(x_n) \end{aligned}$$

(Copiamos)

$$I_{(-\infty, t-\delta]}(x) \leq h(x) \text{ y } h(x_n) \leq I_{(-\infty, t]}(x_n)$$

Tomando ~~probabilidad~~ ^{esperanza}, resulta:

$$\underbrace{P(X \leq t-\delta)}_{F_X(t-\delta)} \leq E(h(x)) \text{ y } E(h(x_n)) \leq \underbrace{P(X_n \leq t)}_{F_{X_n}(t)}$$

Tomando $\liminf_{n \rightarrow \infty}$

$$F_X(t-\delta) \leq E(h(x)) \stackrel{x \text{ Hipótesis}}{=} \lim_{n \rightarrow \infty} E(h(x_n)) \leq \liminf_{n \rightarrow \infty} F_{X_n}(t) \quad (2)$$

$$F_X(t-\delta) \leq E(h(x)) \leq \liminf_{n \rightarrow \infty} F_{X_n}(t)$$

por (2)

Ahora, tomando límite cuando $\delta \rightarrow 0$ y usando que F_X es continua en t tenemos:

$$F_X(t) \leq \liminf_{n \rightarrow \infty} F_{X_n}(t)$$

Combinando las 2 desigualdades resaltadas, tenemos:

$$F_X(t) \leq \liminf F_{X_n}(t) \leq \limsup F_{X_n}(t) \leq F_X(t) \quad \checkmark$$

Esta equivalencia entre la convergencia en distribución y la convergencia de $E(g(X_n))$ a $E(g(X))$ para toda g continua y acotada, tiene varias utilidades, además de garantizar la convergencia de ciertas esperanzas.

- ↳ Permite extender la noción de convergencia en distribución a vectores, sorteando el uso de las funciones de distribución multidimensionales.
- A través de ella pueden probarse algunos de los resultados que ya vimos; por ej que si $X_n \xrightarrow{D} X$ $\Rightarrow h(X_n) \xrightarrow{P} h(Y)$ si h continua.

Corolario 7.24

Sea k un entero fijo. Las funciones g del Teorema 7.23 pueden tomarse en el espacio de funciones acotadas, continuas, derivables con k derivadas continuas y acotadas.

Ley de los Grandes Números: intuición

Si repetimos n veces de manera independiente un mismo experimento, por ejemplo realizamos n mediciones de una cantidad física en un laboratorio (peso, longitud, fluorescencia, etc.), y denotamos por X_i al resultado obtenido en la i -ésima repetición del experimento, entonces para n

suficientemente grande, el promedio $\frac{1}{n} \sum_{i=1}^n X_i$ estará muy cerca de un

número fijo. Intuitivamente, dijimos que este número era la esperanza de X_i . Este razonamiento también le da justificación a la interpretación de las probabilidades como frecuencias relativas en el sentido que

$$P(A) \approx \frac{1}{n} \sum_{i=1}^n I_{\{X_i \in A\}}$$

Es decir, la probabilidad de un evento A está bien aproximada por la frecuencia relativa de las ocurrencias de A en n repeticiones, es decir, la proporción de veces que el resultado del experimento pertenece a A , siempre que realicemos una gran cantidad de experimentos independientes en las mismas condiciones y registremos sus resultados en las variables independientes idénticamente distribuidas (i.i.d.): X_1, \dots, X_n . Nuestro modelo matemático, ¿es capaz de reflejar esta intuición? O, más precisamente, ¿en qué sentido podemos esperar que se produzca la convergencia hacia la esperanza?

Para contestar, vamos a comenzar con un ejemplo de [Ensayos Bernoulli](#).

Recordemos la Fórmula de Stirling ¹

$$n! = \sqrt{n2\pi} n^n e^{-n+\eta(n)}$$

$$\text{con } 0 < \eta(n) < \frac{1}{12n}$$

O, equivalentemente,

$$\lim_{n \rightarrow \infty} \frac{n!}{\sqrt{n2\pi} n^n e^{-n}} = 1. \quad (1)$$

Llamemos

$$b_n = \sqrt{n2\pi} n^n e^{-n}$$

.El límite (1) suele escribirse con la siguiente notación para convergencia de sucesiones: $n! \sim b_n$ cuando $n \rightarrow \infty$ (lo cual indica que el cociente entre ellas tiende a 1 cuando n tiende a ∞).

¹Ver por ejemplo W. Feller. (1968) *An introduction to probability theory and its applications, Vol. I.* o W. Rudin. (1976). *Principles of mathematical analysis.*

Ejemplo 7.6 (fracasamos si somos muy estrictos)

(Ensayos Bernoulli). Sean $(X_i)_{i \geq 1}$ una sucesión de ensayos Bernoulli con probabilidad de éxito $p = 1/2$. Podemos pensar en tiros secuenciales de una moneda equilibrada (o en un paseo al azar simétrico). ¿Cuál es la probabilidad de que la proporción de caras sea exactamente igual a $1/2$? Obviamente, esto solamente es posible si el número de tiros es par. En tal caso, sea $Y \sim Bi(2n, \frac{1}{2})$,

$$P\left(\frac{1}{2n} \sum_{i=1}^{2n} X_i = \frac{1}{2}\right) = P(Y = n) = \binom{2n}{n} 2^{-2n} = a_n$$

Por la fórmula de Stirling obtenemos la siguiente aproximación:

$$\lim_{n \rightarrow \infty} \frac{n!}{b_n} = 1, \text{ luego } \lim_{n \rightarrow \infty} \binom{2n}{n} \frac{(b_n)^2}{b_{2n}} = \lim_{n \rightarrow \infty} d_n = 1, \text{ Entonces}$$

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} d_n \frac{b_{2n}}{(b_n)^2} \frac{1}{2^{2n}} = \lim_{n \rightarrow \infty} d_n \frac{(2n)^{2n} \sqrt{2\pi 2n}}{(n^n \sqrt{2\pi n})^2} \frac{1}{2^{2n}} = \lim_{n \rightarrow \infty} d_n \frac{1}{\sqrt{\pi n}} \xrightarrow{n \rightarrow \infty} 0$$

Ejemplo 7.6, continuación

O sea, $a_n \sim \frac{1}{\sqrt{\pi n}}$.

Esto significa que, para n suficientemente grande, la frecuencia relativa es igual a $1/2$ solamente con una baja probabilidad. Por lo que tiene sentido lo que intuitivamente decimos, “con suerte alrededor de $1/2$ ”.

El ejemplo anterior sugiere que debilitemos nuestra conjectura, como sigue. Para todo $\varepsilon > 0$

$$P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - E(X_1) \right| \leq \varepsilon \right) \xrightarrow{n \rightarrow \infty} 1$$

Es decir, que

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} E(X_1)$$

Este resultado se conoce como la **Ley Débil de los Grandes Números**.

Teorema 7.25 (Ley débil de los grandes números)

Sean $(X_i)_{i \geq 1}$ no correlacionadas, con $E[X_i] = \mu_i$, $V(X_i) = \sigma_i^2$, entonces

Sean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \bar{\mu}_n = E[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mu_i$$

Si $\frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 \rightarrow 0$, entonces

$$\bar{X}_n - \bar{\mu}_n \xrightarrow{P} 0$$

Demostración.

$$P\left(\left|\frac{\sum_{i=1}^n X_i}{n} - \bar{\mu}_n\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2} V\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n \sigma_i^2}{n^2 \varepsilon^2} \rightarrow 0$$

cuando $n \rightarrow \infty$, cualquiera sea $\varepsilon > 0$.



Diremos que las variables $(X_i)_{i \geq 1}$ son i.i.d. para abreviar escribir que son una colección de variables independientes idénticamente distribuidas.

Corolario 7.26 (Ley débil de los grandes números, caso i.i.d.)

Sean $(X_i)_{i \geq 1}$ i.i.d., $E[X_i] = \mu$, $V(X_i) = \sigma^2$, entonces, llamando

$$S_n = \sum_{i=1}^n X_i$$

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow{P} \mu .$$

Veremos más adelante que este mismo resultado vale sin pedir segundo momento.

Ley Fuerte de los Grandes Números

Teorema 7.27 (Ley Fuerte de los Grandes Números)

Sean $(X_i)_{i \geq 1}$ i.i.d., $E[X_i] = \mu$, $V(X_i) = \sigma^2$ (es decir, $E(X_i^2) < \infty$), entonces

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{c.s.} \mu$$

Primer paso: Sin perder generalidad, podemos asumir que $E(X_i) = 0$ pues sino podemos trabajar con $X'_i = X_i - \mu$ que cumple $E(X'_i) = 0$. Luego, asumiendo que tenemos probado el Teo para X'_i veamos cómo probarlo para las X_i :

Sabemos $\frac{1}{n} \sum_{i=1}^n (X_i - \mu) \xrightarrow{c.s.} 0$ o sea, $\frac{1}{n} \sum_{i=1}^n X_i - \mu \xrightarrow{c.s.} 0$

Pero como $g(t) = t + \mu$ es continua, preserva la convergencia cs. ✓

2º-paso: (Asumimos $\mu=0$) Sea $S_n = \sum_{i=1}^n X_i$

Probemos que $\frac{S_n}{n^2} \xrightarrow{\text{c.s.}} 0$. Para ello, sea $\epsilon > 0$.

$$P\left(\left|\frac{S_n}{n} - 0\right| > \epsilon\right) \leq \text{Var}\left(\frac{S_n}{n}\right) \frac{1}{\epsilon^2} = \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n V(X_i) = \frac{n\sigma^2}{\epsilon^2 n^2}$$

↑ Desig de Cheby

$$= \frac{\sigma^2}{\epsilon^2 n}.$$

(pues $E\left(\frac{S_n}{n}\right) = E(X_1) = 0$)

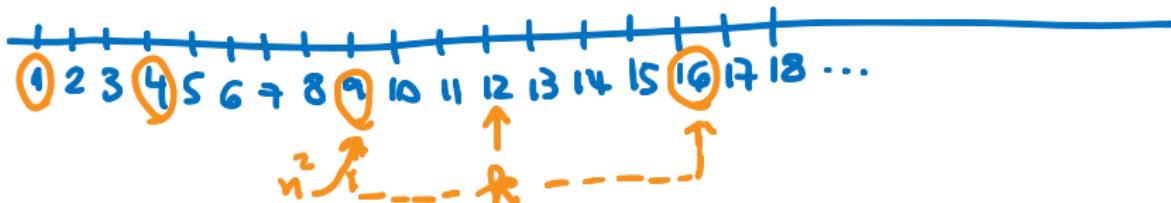
En particular, $\frac{S_n}{n^2}$ cumple que $\sum_{n=1}^{\infty} P\left(\left|\frac{S_n}{n^2} - 0\right| > \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2} \sum_{n=1}^{\infty} \frac{1}{n^2} < +\infty$

⇒ por el criterio de convergencia c.s. que se deduce de Borel - Cantelli, tenemos $\frac{S_n}{n^2} \xrightarrow{\text{c.s.}} E(X_1) = 0$

3er paso: Tenemos una subsucesión que converge c.s.

Queremos probar que toda la sucesión converge.

Sea $k \in \mathbb{N}$. Buscamos los elementos de la subsucesión convergente que están más cerca de k :



Sean n : $n^2 \leq k < (n+1)^2$. Para identificarlo, podemos llamarlo $n_k = \max \{n \in \mathbb{N} : n^2 \leq k\}$.

Queremos probar que $\frac{s_k}{k} \xrightarrow{\text{c.s.}} 0$

$$\left| \frac{s_k}{k} \right| = \left| \frac{s_k - s_{n_k^2}}{k} + \frac{s_{n_k^2}}{k} \right| \leq \left| \frac{s_k - s_{n_k^2}}{k} \right| + \left| \frac{s_{n_k^2}}{k} \right|$$

(copiamos)

$$\left| \frac{S_k}{k} \right| = \left| \frac{S_k - S_{n_k^2}}{k} + \frac{S_{n_k^2}}{k} \right| \leq \left| \frac{S_k - S_{n_k^2}}{k} \right| + \left| \frac{S_{n_k^2}}{k} \right| \leq$$

como $n_k^2 \leq k < (n_k + 1)^2 \Rightarrow \frac{1}{k} \leq \frac{1}{n_k^2}$

$$\leq \left| \frac{S_k - S_{n_k^2}}{k} \right| + \left| \frac{S_{n_k^2}}{n_k^2} \right| \xrightarrow{\text{c.s.}} 0$$

Basta ver que $w_k = \frac{S_k - S_{n_k^2}}{k} \xrightarrow{\text{c.s.}} 0$. Sea $\varepsilon > 0$.

$$P(|w_k| > \varepsilon) = P\left(\left|\sum_{j=n_k^2+1}^k x_j\right| > \varepsilon k\right) \leq \frac{\sqrt{\left(\sum_{j=n_k^2+1}^k x_j\right)}}{\varepsilon^2 k^2} = \frac{\sigma \sqrt{k - n_k^2}}{\varepsilon^2 k^2}$$

Comentario:

Podemos usarlo q. $E(w_k) = E(x_j) = 0$

$$\text{Tenemos: } P\left(\left|\frac{S_k - S_{n_k^2}}{k}\right| > \varepsilon\right) \leq \frac{\sigma^2(k - n_k^2)}{\varepsilon^2 k^2} \stackrel{?}{\leq} \frac{\sigma^2 2\sqrt{k}}{\varepsilon^2 k^2} = \frac{c}{k^{3/2}}$$

como $n_k^2 \leq k < (n_k + 1)^2 \Rightarrow 0 \leq k - n_k^2 < (n_k + 1)^2 - n_k^2 = 2n_k + 1$

$$\Rightarrow 0 \leq k - n_k^2 \leq 2n_k \leq 2\sqrt{k}$$

y también $n_k \leq \sqrt{k}$

es una
serie
sumable

Luego, x el criterio de conv. c.s. sabemos que $\frac{S_k - S_{n_k^2}}{k} \xrightarrow{\text{c.s.}} 0$

$$\Rightarrow \left| \frac{S_k - S_{n_k^2}}{k} \right| \rightarrow 0 . \text{ Finalmente, tenemos}$$

$$\left| \frac{S_k}{k} \right| \leq \left| \frac{S_k - S_{n_k^2}}{k} \right| + \left| \frac{S_{n_k^2}}{n_k^2} \right| \xrightarrow{\text{c.s.}} 0 \text{ porque suma de sucesiones de v.a que conv c.s. a } 0, \text{ tienden a cero.}$$

$\xrightarrow{\text{c.s.}} 0$ $\xrightarrow{\text{c.s.}} 0$

Dos comentarios

1) La LGN, versión fuerte sale con el supuesto de $E|X_i| < \infty$. La prueba puede verse en Durrett, (2019) Probability Theory and Examples, x-ejemplo.

2) Es un ejercicio de la práctica probar :

Teo: Si $(X_i)_{i \in \mathbb{N}}$ son v.a.iid con $E(|X_i|) = +\infty$ entonces $\limsup_{n \rightarrow \infty} \frac{S_n}{n} = +\infty$ cs.

Lo cual prueba que la condición de esperanza finita es necesaria para que valga la LGN.

Aplicaciones

1) Teorema de Weierstrass: Los polinomios son densos en el espacio $C([0,1]) = \{ f : [0,1] \rightarrow \mathbb{R} \text{ continuas} \}$ con la norma del supremo, $\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|$

Damos una demostración constructiva.

Sea $f \in C[0,1]$, definimos

$$q_n(x) = \sum_{k=0}^n f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}, \quad x \in [0,1]$$

El polinomio
de Bernstein de
grado n .

Sean $y_i \sim Be(x)$ iid, $1 \leq i \leq n$.

Entonces $S_n = \sum_{i=1}^n y_i \sim Bi(n, x)$

$$E(f(S_n)) = \sum_{k=0}^n f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k} = q_n(x)$$

Logramos escribir a lo que nos interesa, $q_n(x)$, como la esperanza de una r.v. Por la ley débil de los Grandes Números, tenemos que

$$\frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{P} E(Y_1) = x$$

f es continua
continua

$\Rightarrow f\left(\frac{S_n}{n}\right) \xrightarrow{P} f(x)$. Como f es continua en $[0,1]$ entonces es acotada,

$$|f(t)| \leq \|f\|_\infty \quad \forall t \in [0,1]$$

\Rightarrow

Prop 7.7
(teo de la convergencia acotada)

$f\left(\frac{S_n}{n}\right) \xrightarrow{L^1} f(x) \Rightarrow q_n(x) = E\left(f\left(\frac{S_n}{n}\right)\right) \xrightarrow{\text{Lema 7.8}} E(f(x))$

$f(x)$

Como esta convergencia sucede para cada $x \in [0,1]$, acabamos de probar que $q_n(x)$ convergen puntualmente a $f(x)$.

Veamos la convergencia uniforme. Sea $\varepsilon > 0$.

Como f es uniformemente continua en $[0,1]$ $\exists \delta > 0 /$
si $|x-y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon \quad \forall x, y \in [0,1]$.

Queremos acotar:

$$|q_n(x) - f(x)| = \left| E(f(s_{y_n}) - f(x)) \right| \leq E |f(\frac{s_n}{n}) - f(x)|$$

Luego,

$$\left| f\left(\frac{s_n}{n}\right) - f(x) \right| = \left| f\left(\frac{s_n}{n}\right) - f(x) \right| \cdot I_A + \left| f\left(\frac{s_n}{n}\right) - f(x) \right| I_{A^c}$$

donde $A = \left\{ \left| \frac{s_n}{n} - x \right| < \delta \right\}$

Entonces, $\left| f\left(\frac{s_n}{n}\right) - f(x) \right| \leq \varepsilon I_A + 2 \|f\|_\infty I_{A^c}$

$$\Rightarrow E\left(\left| f\left(\frac{s_n}{n}\right) - f(x) \right|\right) \leq \varepsilon P(A) + 2 \|f\|_\infty P(A^c) \stackrel{E\left(\frac{s_n}{n}\right) = x}{\leq} \varepsilon + 2 \|f\|_\infty \text{Var}\left(\frac{s_n}{n}\right) \cdot \frac{1}{\delta^2}$$

cheby.

$$(\text{Copiamos}) \quad E\left(\left|f\left(\frac{S_n}{n}\right) - f(x)\right|\right) \leq \varepsilon + \frac{2\|f\|_\infty}{\delta^2} \text{Var}\left(\frac{S_n}{n}\right)$$

$$= \varepsilon + \frac{2\|f\|_\infty}{\delta^2 n^2} \underbrace{n \times (1-x)}_{\text{Var}(S_n), S_n \sim Bi(n, x)} \leq \varepsilon + \frac{2\|f\|_\infty}{\delta^2 n}$$

$$\Rightarrow |q_n(x) - f(x)| \leq \varepsilon + \frac{2\|f\|_\infty}{\delta^2 n} \quad \forall x \in [0,1]$$

$$\sup_{x \in [0,1]} |q_n(x) - f(x)| \leq \varepsilon + \frac{2\|f\|_\infty}{\delta^2 n}$$

$\rightarrow \limsup_{n \rightarrow \infty} \|q_n - f\|_\infty \leq \varepsilon + \varepsilon > 0$ finalmente, resulta

$\lim_{n \rightarrow \infty} \|q_n - f\|_\infty = 0$, probando el resultado. ✓

A partir de esta prueba, es fácil ver que si f es monótona, los polinomios de Bernstein también lo son.

Observemos que si $U_i \sim U(0,1)$ ^{indep} y definimos

$$Y_i = I_{[0,x]}(U_i) \text{ tenemos } Y_i \sim Be(x) \text{ iid}$$

Supongamos f creciente, y tenemos $x_1 < x_2$

como $I_{[0,x_1]}(t) \leq I_{[0,x_2]}(t)$



$$\Rightarrow \frac{1}{n} \sum_{i=1}^n I_{[0,x_1]}(U_i) \leq \frac{1}{n} \sum_{i=1}^n I_{[0,x_2]}(U_i) \rightarrow \text{"Acoplamiento!!"}$$

y, x0 tanto: $q_n(x_1) = E\left(\frac{1}{n} \sum_{i=1}^n I_{[0,x_1]}(U_i)\right) \leq E\left(\frac{1}{n} \sum_{i=1}^n I_{[0,x_2]}(U_i)\right) = q_n(x_2)$

Este gráfico muestra 2 ejemplos de funciones f y 4 polinomios de Bernstein (p_{2k} , $k=0, 1, \dots, 4$)
(fuente: Georgii, H.O. stochastics: introduction to probability & Statistics)

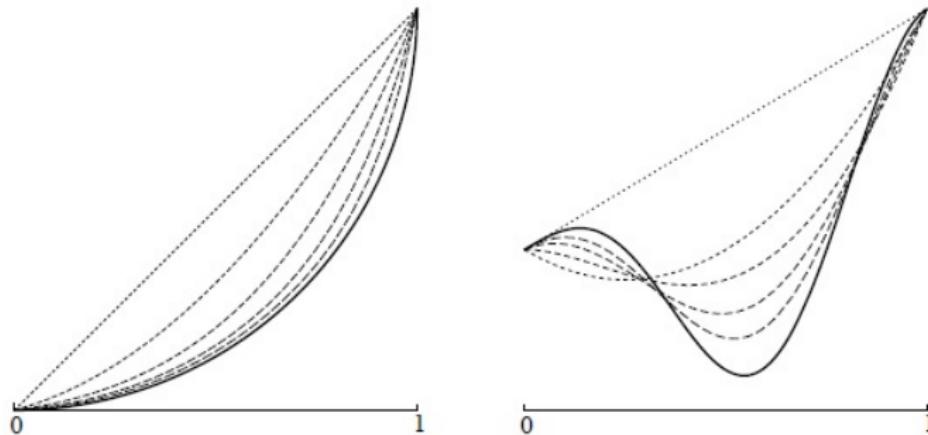


Figure 5.2. The Bernstein polynomials f_{2k} for $k = 0, \dots, 4$ (dashed) for two different functions f (solid lines).

Aplicación de la Ley de los Grandes Números

Segundo ejemplo: calcular integrales definidas aproximadamente (Monte Carlo)

Queremos calcular una integral definida, $\int_0^1 g(x)dx$ pero la integración no puede hacerse usando primitivas elementales. La forma más común de hallarla es usando métodos numéricos en los cuales la integral es aproximada por una suma. Otra posibilidad se denomina método de Monte Carlo y funciona del siguiente modo.

Sean X_1, \dots, X_n v.a.i.i.d., $X_i \sim \mathcal{U}(0, 1)$. Entonces:

① $E(g(X_1)) = \int_{-\infty}^{+\infty} g(t)f_{X_1}(t)dt = \int_0^1 g(t)dt.$

② La Ley de los Grandes Números nos asegura que

$$\frac{1}{n} \sum_{i=1}^n g(X_i) \xrightarrow{P} \int_0^1 g(t)dt. \text{ Más aún,}$$

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n g(X_i) - \int_0^1 g(t)dt\right| \geq \varepsilon\right) \leq \frac{V(g(X_1))}{n\varepsilon^2} = \frac{c}{n\varepsilon^2}$$

El método de Monte Carlo para calcular (o aproximar) integrales consiste en elegir X_1, \dots, X_n realizaciones independientes de la variable aleatoria con distribución uniforme, tomando *n bien grande*, y luego aproximar

$$\int_0^1 g(t)dt \text{ por } \frac{1}{n} \sum_{i=1}^n g(X_i(\omega)).$$

Así, con una alta probabilidad se obtendrá una buena aproximación del valor de la integral.

Este método puede adaptarse cuando la integral de interés está definida en un intervalo $[a, b]$ arbitrario o en un conjunto acotado de \mathbb{R}^k (**¿cómo?**)

El método de Monte Carlo no es especialmente eficiente para $k = 1$ (integrales de una variable) pero la eficiencia mejora al aumentar el k .

Como ejemplo, aproximemos la integral $I = \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$, con

$n = 1000$. El valor verdadero es $I = 0,3413447$. Lo que hacemos está resumido en el siguiente

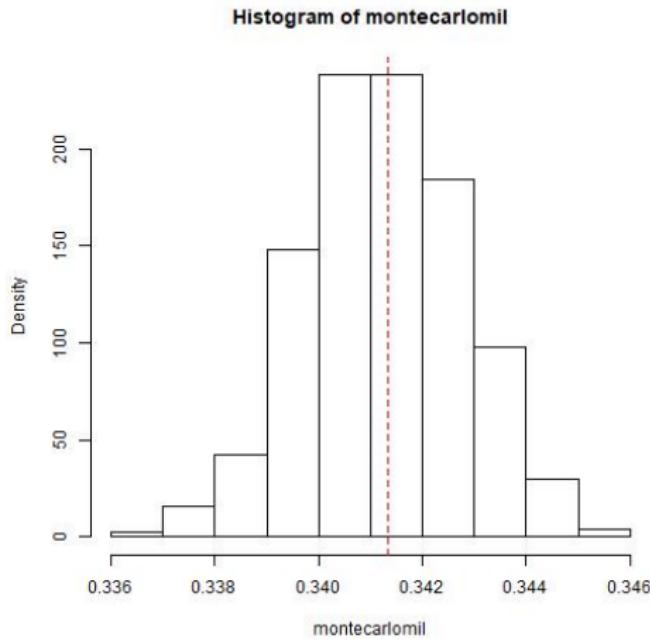
ALGORITMO DE MONTE CARLO

- ① pedirle a la computadora que genere n realizaciones de una uniforme $(0,1)$,
- ② las transformamos, aplicándoles la función $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$,
- ③ y luego promediarlas.

El problema con esto es que cada vez que repitamos el algoritmo obtendremos resultados distintos. Para darnos una idea de cuán distintos son estos resultados, repetimos 500 veces el procedimiento, y comparemos los resultados. Le pedimos al R que repita el algoritmo anterior, y guardamos los 500 resultados en un vector, que se llama **montecarlomil**. ¿Cómo “vemos” 500 resultados? Podemos hacer un histograma.

Script de R: **montecarlopaintegrales.R**

Figura 1: Histograma de $B = 500$ replicaciones. Cada una es un promedio de $n = 1000$ observaciones independientes, $\frac{1}{n} \sum_{i=1}^n g(X_i(\omega))$. La línea vertical roja representa la integral verdadera que estamos queriendo aproximar por el método de Montecarlo.



A todo lo realizado suele llamársele una simulación. ¿Qué es una simulación?

SIMULACIÓN (en este ejemplo)

- Fijamos $n = 1000$, la cantidad de observaciones promediadas.
- Fijamos $B = 500$, la cantidad de repeticiones del algoritmo.
- Repetimos el algoritmo descripto antes, de forma independiente, B veces, cada vez usando n observaciones. Guardamos los B resultados obtenidos, usualmente en un vector.
- Resumimos el vector obtenido, de forma gráfica, por ejemplo, con un histograma.

Por supuesto, por el azar involucrado, cada vez que repitamos la simulación obtendremos resultados diferentes. ¿Cuánto mejoramos cuando aumentamos la cantidad de observaciones que generamos y promediamos? Para compararlas, repitamos la simulación tomando,

- $n = 1,000$ (ya lo hicimos)
- $n = 10,000$
- $n = 100,000$

Para entender los resultados obtenidos, hacemos un gráfico con los histogramas de las 3 simulaciones de colores distintos.

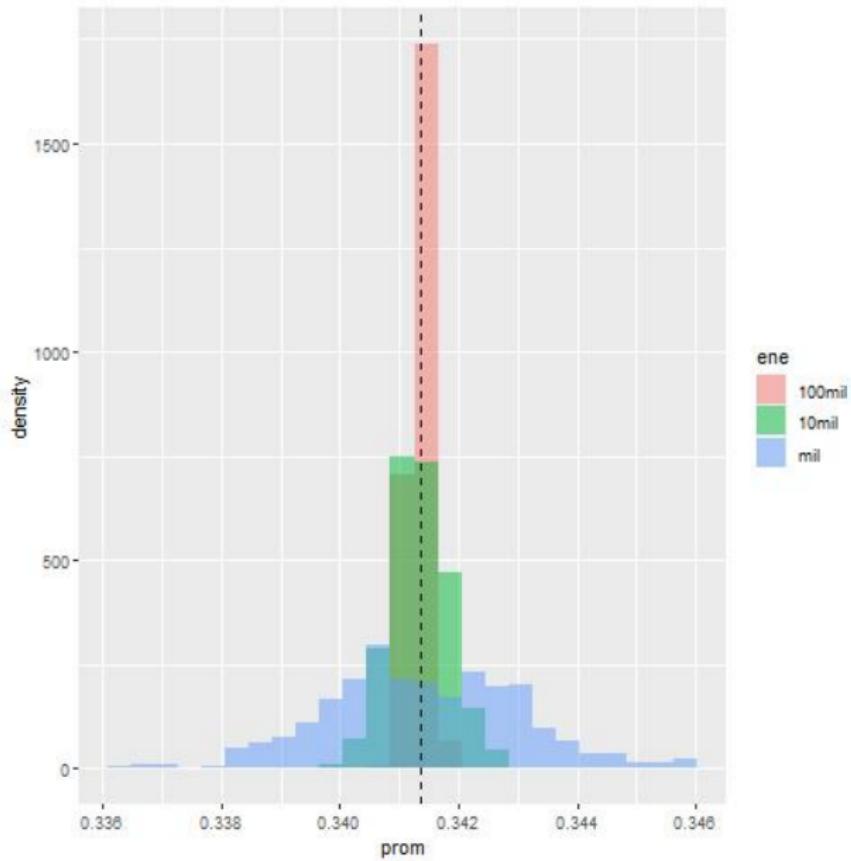


Figura 2: Histograma de las 3 simulaciones, cada una basada en $B = 500$ replicaciones, con $n = 1,000$ en celeste, $n = 10,000$ en verde y $n = 100,000$ en rosa. Vemos cómo al aumentar el n se gana precisión, ya que se concentran alrededor de la línea vertical punteada, graficada sobre el valor verdadero. Esto es consecuencia de la LGN.

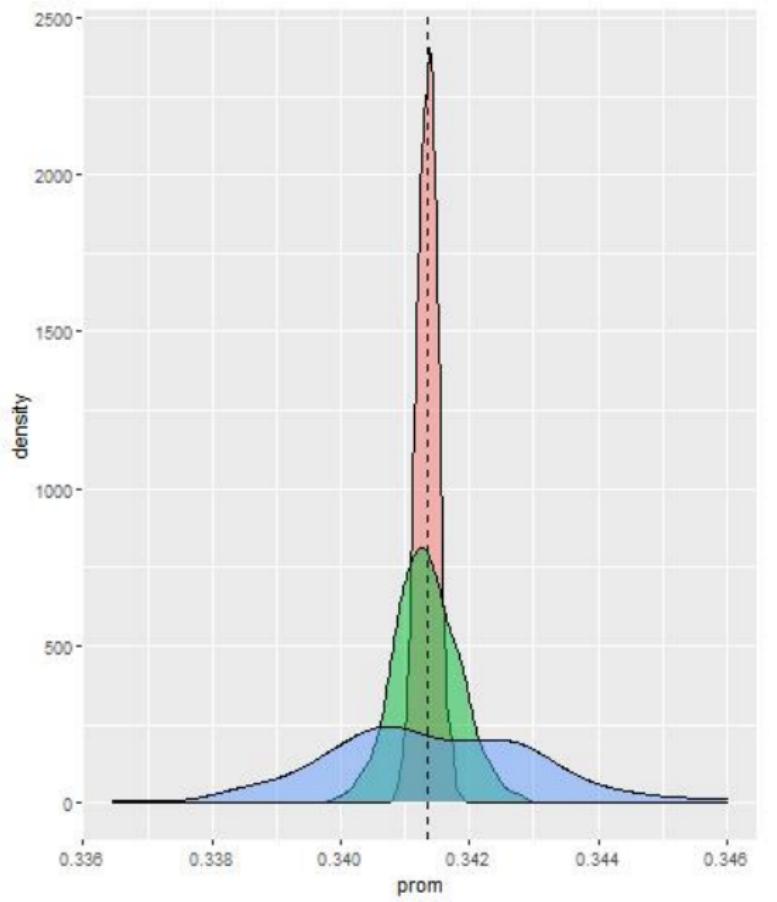


Figura 3: Alternativa al histograma de las 3 simulaciones: la curva representa una aproximación a la densidad de las observaciones, una versión “continua” del histograma que se denomina estimación no paramétrica de la densidad.

Aplicación de la Ley de los Grandes Números

Tercer ejemplo: aproximar probabilidades mediante la frecuencia relativa

Si queremos aproximar la probabilidad de que ocurra un evento y somos capaces de reproducir un experimento en condiciones idénticas e independientes, podemos generar X_1, \dots, X_n variables aleatorias i.i.d. Sea A el evento de interés. Luego, definimos las indicadoras de que el resultado del experimento pertenece al conjunto A :

$$I_A(X_i) = \begin{cases} 1 & \text{si } X_i \in A \\ 0 & \text{caso contrario} \end{cases}$$

Entonces, $Y_i = I_A(X_i) \sim Be(P(A))$ i.i.d. y además

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n I_A(X_i) = \frac{1}{n} \# \{i : X_i \in A\} = \text{frecuencia relativa de la ocurrencia de } A$$

Por la Ley de los Grandes Números, sabemos que

$\bar{Y}_n \xrightarrow{c.s.} E(Y_1) = P(X_1 \in A)$ que es la probabilidad que queremos hallar.

Veamos dos ejemplos que ya calculamos.

- ① La probabilidad de que la suma de dos dados equilibrados dé 8.
- ② La probabilidad de que en un grupo de $r = 23$ personas no haya dos o más que cumplan el mismo día.

Script de R: generar2020.R

8. Teorema Central del Límite

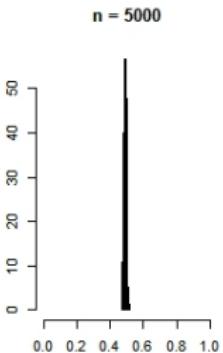
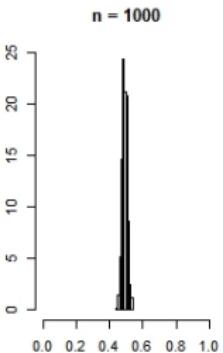
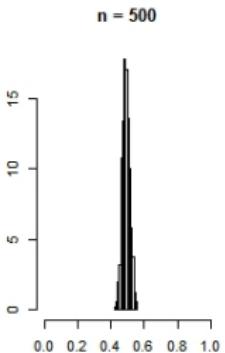
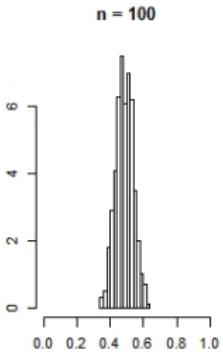
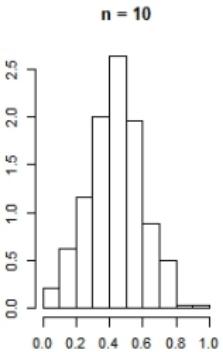
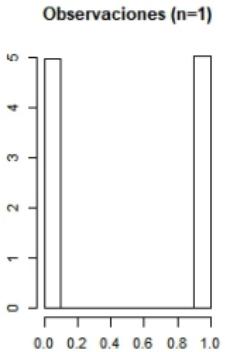
Probabilidades y Estadística (M)

María Eugenia Szretter Noste

Departamento de Matemática e
Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Primer cuatrimestre 2020



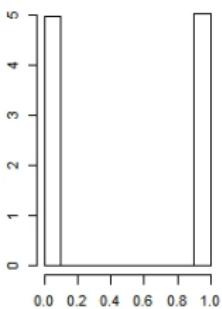


Histogramas del final del tercer ejemplo de la LGN, graficamos el histograma de \bar{X}_n , con $n = 1, \dots, 5000$, cada uno basados en 500 replicaciones del experimento de elegir 23 personas al azar, éxito es que no coincida ninguna fecha de cumpleaños.

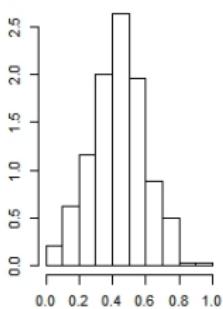
Mantenemos el ejex entre 0 y 1. Vemos la LGN en acción: cuando n crece, los promedios se concentran alrededor de la verdadera probabilidad, que es

0.4927. ▶◀ ⏷ ⏸ ⏹ 🔍 2

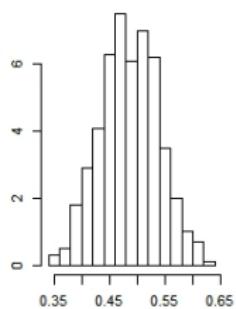
Observaciones (n=1)



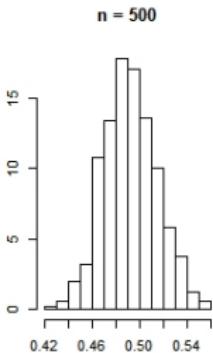
n = 10



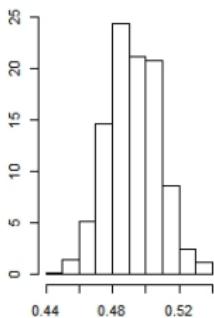
n = 100



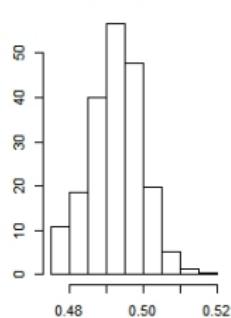
n = 500



n = 1000



n = 5000



Mismos histogramas que en la página anterior, graficamos el histograma de \bar{X}_n con $n = 1, \dots, 5000$, cada uno basados en 500 replicaciones del experimento de elegir 23 personas al azar.

Dejamos que el R acomode la escala en ambos ejes. Acá **no vemos** la concentración tan fácilmente, ya que las escalas se están corrigiendo. Pero sí vemos que los histogramas (apropiadamente reescalados) se parecen a los de la normal.

¿Magia? No, TCL

Herramientas previas

Vamos a usar una serie de resultados que ya vimos.

Teorema 7.23 (caracterización de la conv en distrib vía esperanzas)

Son equivalentes:

- a) $X_n \xrightarrow{\mathcal{D}} X$
- b) $E[g(X_n)] \rightarrow E[g(X)]$ para toda g función continua y acotada.

Y su corolario.

Corolario 7.24

Sea k un entero fijo. Las funciones g del Teorema 7.23 pueden tomarse en el espacio de funciones acotadas, continuas, derivables con k derivadas continuas y acotadas.

Herramientas previas

- ① Si $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ son independientes entonces $aX + bY + c \sim \mathcal{N}(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2)$ para todo $a, b, c \in \mathbb{R}$ con $a \neq 0$ ó $b \neq 0$ (ejercicio de la práctica).
- ② En particular, si $X_1, X_2 \sim \mathcal{N}(\mu, \sigma^2)$ son independientes, entonces $X_1 + X_2 \sim \mathcal{N}(2\mu, 2\sigma^2)$

Transformación. Cabe la siguiente pregunta, dada una variable aleatoria S cualquiera en \mathcal{L}^2 , ¿qué transformación (que la modifique lo menos posible) puedo hacerle para que tenga esperanza cero, y varianza 1? Es decir, ¿puedo elegir reales a y b para que $Z = b(S + a)$ tenga esperanza 0 y varianza 1?

Respuesta: $a = -E(S)$ y $b = \frac{1}{\sqrt{V(S)}}$. A esta transformación, que convierte a la variable S en Z se la conoce como **Estandarización de la variable S** .

Luego,

$$\frac{S - E(S)}{\sqrt{V(S)}}$$

tiene esperanza 0 y varianza 1.

Sean $(X_i)_{i \geq 1}$ variables aleatorias i.i.d. con $E[X_1] = \mu$ y $V(X_1) = \sigma^2$.

- Calculemos esperanza y varianza de la suma de n v.a.i.i.d. entonces tenemos que, $S_n = \sum_{i=1}^n X_i$,

$$E\left(\sum_{i=1}^n X_i\right) = nE[X_1] = n\mu$$

$$V\left(\sum_{i=1}^n X_i\right) = nV(X_1) = n\sigma^2$$

Luego, $S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$ tiene esperanza 0 y varianza 1.

- Calculemos esperanza y varianza del promedio de n v.a.i.i.d. entonces tenemos que, $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = \frac{S_n}{n}$

$$E(\bar{X}_n) = E[X_1] = \mu$$

$$V(\bar{X}_n) = \frac{1}{n} V(X_1) = \frac{\sigma^2}{n}$$

Luego, $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ tiene esperanza 0 y varianza 1.

¿Y si agregamos el supuesto de normalidad?

Sean $(X_i)_{i \geq 1}$ variables aleatorias i.i.d. con $E[X_1] = \mu$ y $V(X_1) = \sigma^2$. Y además suponemos que tienen distribución normal. Entonces $X_i \sim \mathcal{N}(\mu, \sigma^2)$ independientes. Luego

$$S_n = \sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Leftrightarrow S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0, 1).$$

O, equivalentemente, escrito en términos del promedio,

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

El Teorema Central del Límite nos dice que si en lugar de variables normales trabajamos con suma de variables independientes e identicamente distribuidas (i.i.d.) con esperanza μ y varianza σ^2 , la suma de ellas se comporta (**aprox**) como si estuviéramos sumando **variables normales**.

Teorema 8.1 (Teorema Central del Límite)

Sean $(X_i)_{i \geq 1}$ variables aleatorias i.i.d. con $E[X_1] = \mu$ y $V(X_1) = \sigma^2 < +\infty$, entonces tenemos que

$$\frac{\sum_{i=1}^n X_i - E(\sum_{i=1}^n X_i)}{\sqrt{V(\sum_{i=1}^n X_i)}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{\mathcal{D}} Z \quad Z \sim \mathcal{N}(0, 1),$$

equivalentemente

$$\frac{\sqrt{n}}{\sigma} \left\{ \frac{\sum_{i=1}^n X_i}{n} - \mu \right\} \xrightarrow{\mathcal{D}} Z \quad Z \sim \mathcal{N}(0, 1).$$

Es decir, que si Φ es la función de distribución acumulada de una $Z \sim \mathcal{N}(0, 1)$, $\Phi(t) = P(Z \leq t)$, entonces

$$P \left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \leq t \right) \xrightarrow{n \rightarrow \infty} \Phi(t),$$

para todo $t \in \mathbb{R}$.

Hay tradicionalmente dos pruebas para el Teorema Central del Límite (o TCL). La más difundida usa la función característica, que definiremos en breve. Ahora daremos una demostración debida a Lindeberg, que es más elemental ya que no requiere estas herramientas. Pero la idea básica en ambos argumentos es estimar el valor esperado de una función suave de una suma de variables independientes usando el desarrollo de Taylor y acotando el error.

Demostración del TCL.

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \Leftrightarrow \frac{\sum_{i=1}^n \frac{X_i - E(X_1)}{\sqrt{V(X_1)}}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Si llamamos $Y_i = \frac{X_i - E(X_1)}{\sqrt{V(X_1)}}$ entonces $E[Y_i] = 0$ y $V(Y_i) = 1$ y el enunciado se reduce a

$$\frac{\sum_{i=1}^n Y_i}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Supongamos entonces que las X_i son tales que $E[X_i] = 0$ y $V(X_i) = 1$.

Por el Teorema 7.23 y su Corolario 7.24, basta ver que

$E\left[f\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right)\right] \rightarrow E[f(Z)]$ para toda $f : \mathbb{R} \rightarrow \mathbb{R}$ función acotada, continua, derivable con $k = 3$ derivadas continuas y acotadas.

Construimos una sucesión $\{Z_i\}_{i \geq 1}$ de variables aleatorias i.i.d. con distribución $\mathcal{N}(0, 1)$, independiente de las $\{X_i\}_{i \geq 1}$. Para todo n ,

$$\frac{\sum_{i=1}^n Z_i}{\sqrt{n}} \sim \mathcal{N}(0, 1)$$

. Por el Teorema 7.23 + Coro 7.24 basta ver que

$$E\left[f\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right) - f\left(\frac{\sum_{i=1}^n Z_i}{\sqrt{n}}\right)\right] \xrightarrow{n \rightarrow \infty} 0$$

La clave es expresar a $f\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right) - f\left(\frac{\sum_{i=1}^n Z_i}{\sqrt{n}}\right)$ como una suma telescópica .

Sean

$$S := S_0 := (X_1 + X_2 + X_3 + \cdots + X_n) \frac{1}{\sqrt{n}}$$

$$S_1 := (\textcolor{orange}{Z}_1 + X_2 + X_3 + \cdots + X_n) \frac{1}{\sqrt{n}}$$

$$S_2 := (\textcolor{orange}{Z}_1 + \textcolor{orange}{Z}_2 + X_3 + \cdots + X_n) \frac{1}{\sqrt{n}}$$

⋮ ⋮ ⋮

$$T := S_n := (\textcolor{orange}{Z}_1 + \textcolor{orange}{Z}_2 + \textcolor{orange}{Z}_3 + \cdots + \textcolor{orange}{Z}_n) \frac{1}{\sqrt{n}}$$

Queremos mostrar que la distribución de S está cerca de la de T que es $\mathcal{N}(0, 1)$, es decir, que $E(f(S))$ está cerca de la $E(f(T))$ para toda $f \in C^3(-\infty, \infty)$, tales que tanto ellas como sus derivadas $f^{(i)}, i = 0, 1, 2, 3$ están uniformemente acotadas por K . Claramente,

$$|E[f(S)] - E[f(T)]| \leq \sum_{j=1}^n |E[f(S_j)] - E[f(S_{j-1})]|$$

Comparemos entonces $f(S_j)$ con $f(S_{j-1})$. Llamemos R_j a la suma de los términos comunes entre S_{j-1} y S_j .

O sea

$$R_j = (Z_1 + Z_2 + Z_3 + \cdots + Z_{j-1} + X_{j+1} + \cdots + X_n) \frac{1}{\sqrt{n}},$$

Entonces $S_{j-1} = R_j + X_j \frac{1}{\sqrt{n}}$ y $S_j = R_j + Z_j \frac{1}{\sqrt{n}}$.

Observemos que por construcción R_j y X_j son independientes, y también lo son R_j y Z_j .

Necesitamos comparar $E[f(S_{j-1})] = E\left[f\left(R_j + \frac{X_j}{\sqrt{n}}\right)\right]$ y

$E[f(S_j)] = E\left[f\left(R_j + \frac{Z_j}{\sqrt{n}}\right)\right].$

Necesitamos comparar $E\left[f\left(R_j + \frac{X_j}{\sqrt{n}}\right)\right]$ y $E\left[f\left(R_j + \frac{Z_j}{\sqrt{n}}\right)\right]$.

Consideraremos el desarrollo de Taylor de f alrededor de x_0 hasta el tercer término

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2!}f^{(2)}(x_0)(x - x_0)^2 + \frac{1}{3!}f^{(3)}(\xi)(x - x_0)^3$$

donde $\xi \in (x_0, x)$. Llamemos $T(x) = \frac{1}{3!}f^{(3)}(\xi)$, entonces

- $T(x)$ está acotada por $\frac{K_0}{3!} = K$

- $\lim_{x \rightarrow x_0} \frac{T(x)(x - x_0)^3}{(x - x_0)^2} = 0$

Tomando $x = R_j + \frac{X_j}{\sqrt{n}}$ y $x_0 = R_j$ obtenemos

$$f\left(R_j + \frac{X_j}{\sqrt{n}}\right) = f(R_j) + \frac{X_j}{\sqrt{n}}f'(R_j) + \left(\frac{X_j}{\sqrt{n}}\right)^2 \frac{1}{2!}f^{(2)}(R_j) + \left(\frac{X_j}{\sqrt{n}}\right)^3 T(R_j + \frac{X_j}{\sqrt{n}})$$

Lo mismo para $R_j + \frac{Z_j}{\sqrt{n}}$:

$$f\left(R_j + \frac{Z_j}{\sqrt{n}}\right) = f(R_j) + \frac{Z_j}{\sqrt{n}}f'(R_j) + \left(\frac{Z_j}{\sqrt{n}}\right)^2 \frac{1}{2!}f^{(2)}(R_j) + \left(\frac{Z_j}{\sqrt{n}}\right)^3 T(R_j + \frac{Z_j}{\sqrt{n}})$$

Entonces, asumiendo que las X 's tienen tercer momento, podemos tomar esperanza en cada una de estas dos identidades y restar las ecuaciones resultantes. **Observación:** Hacemos este supuesto para simplificar la demostración. Si las X 's no tuvieran tercer momento finito, la demostración puede hacerse igual trabajando más con la expresión del resto, ver el libro Georgii, H.O. *Stochastics, Introduction to Probability and Statistics*.

(copiamos)

$$f\left(R_j + \frac{Z_j}{\sqrt{n}}\right) = f(R_j) + \frac{Z_j}{\sqrt{n}} f'(R_j) + \left(\frac{Z_j}{\sqrt{n}}\right)^2 \frac{1}{2!} f^{(2)}(R_j) + \left(\frac{Z_j}{\sqrt{n}}\right)^3 T\left(R_j + \frac{Z_j}{\sqrt{n}}\right)$$

Resulta:

1. Como $E(X_j) = E(Z_j) = 0$, y por la independencia de X_j, R_j , y de Z_j, R_j , tenemos $E(X_j f'(R_j)) = E(X_j) E(f'(R_j)) = 0 = E(Z_j f'(R_j))$
2. Como $V(X_j) = V(Z_j)$, y volviendo a usar la independencia, tenemos $E(X_j^2 f^{(2)}(R_j)) = E(Z_j^2 f^{(2)}(R_j))$.

Luego, los términos del primer y segundo orden se cancelan.

Nos queda entonces:

$$\begin{aligned} E[f(S_j)] - E[f(S_{j-1})] &= E\left[f\left(R_j + \frac{X_j}{\sqrt{n}}\right)\right] - E\left[f\left(R_j + \frac{Z_j}{\sqrt{n}}\right)\right] \\ &= E\left[\frac{X_j^3}{n^{3/2}} T\left(\textcolor{red}{R}_j + \frac{X_j}{\sqrt{n}}\right)\right] - E\left[\frac{Z_j^3}{n^{3/2}} T\left(\textcolor{red}{R}_j + \frac{Z_j}{\sqrt{n}}\right)\right] \end{aligned}$$

Como T está acotada por K ,

$$\begin{aligned} |E[f(S_j)] - E[f(S_{j-1})]| &\leq E\left|\frac{X_j^3}{n^{3/2}} T\left(\textcolor{red}{R}_j + \frac{X_j}{\sqrt{n}}\right)\right| + E\left|\frac{Z_j^3}{n^{3/2}} T\left(\textcolor{red}{R}_j + \frac{Z_j}{\sqrt{n}}\right)\right| \\ &\leq \frac{K}{n^{3/2}} \left(E|X_j|^3 + E|Z_j|^3\right) \leq \frac{K_2}{n^{3/2}} \end{aligned}$$

Entonces

$$|E[f(S)] - E[f(T)]| \leq \sum_{j=1}^n |E[f(S_j)] - E[f(S_{j-1})]| \leq n \frac{K_2}{n^{3/2}} = \frac{K_2}{\sqrt{n}}$$

Como $\frac{K_2}{\sqrt{n}} \xrightarrow{n \rightarrow \infty} 0$, hemos probado el resultado.

Comentarios del TCL

- 1. Interpretación del TCL:** las probabilidades de eventos que involucran el promedio o suma de v.a.i.i.d pueden ser aproximadas usando una distribución normal. Son los cálculos de probabilidades los que pueden ser aproximados, no las variables *per se*.
- 2. ¿Por qué** aparece la distribución $\mathcal{N}(0, 1)$ como distribución límite? Esto vale por la siguiente propiedad de estabilidad de la distribución normal: si las $\{X_i\}_{i \geq 1}$ son variables aleatorias i.i.d. con distribución $\mathcal{N}(0, 1)$, entonces, para un n arbitrario, $S_n^* = \frac{\sum_{i=1}^n X_i - E(S_n)}{\sqrt{V(S_n)}}$ también tiene distribución $\mathcal{N}(0, 1)$. Más aún, la distribución $\mathcal{N}(0, 1)$ es la única distribución de probabilidad en \mathbb{R} con varianza finita que cumple esa propiedad. Puesto que si las $\{X_i\}_{i \geq 1}$ fueran variables aleatorias i.i.d. con distribución G que satisficiera dicha propiedad de estabilidad, tendríamos que S_n^* también tendría distribución G y,

$$S_n^* \xrightarrow{\mathcal{D}} X \quad \text{con } X \sim G \text{ y por el TCL, } G = \mathcal{N}(0, 1).$$

Comentarios del TCL

3. Sean $(X_i)_{i \geq 1}$ variables aleatorias i.i.d. con $E[X_1] = \mu$ y $V(X_1) = \sigma^2$. Si el tercer momento de las X_i existe (i.e., $X_i \in \mathcal{L}^3$), podemos acotar la tasa de convergencia como sigue:

$$\sup_{t \in \mathbb{R}} |F_{S_n^*}(t) - \Phi(t)| \leq 0,8 \frac{E(|X_1 - E(X_1)|^3)}{\sigma^3} \frac{1}{\sqrt{n}}$$

con $S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma^2}$. Este es el **Teorema de Berry-Esséen**. Se puede encontrar una prueba en

- R. Durrett. Probability: Theory and Examples, 2019. Versión disponible online en su página web <https://services.math.duke.edu/~rtd/>.
- A. N. Shirayev. Probability. Springer, New York etc., 1984.

Este resultado nos da una cota para el error que cometemos al reemplazar a $S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma^2}$ por una variable aleatoria $Z \sim \mathcal{N}(0, 1)$.

4. La LGN y el TCL son compatibles. La LGN nos dice que

$$\bar{X}_n \xrightarrow{c.s.} E(X_1)$$

Equivalentemente,

$$\bar{X}_n - E(X_1) \xrightarrow{c.s.} 0,$$

que es un número fijo. O sea que $\bar{X}_n - E(X_1) \xrightarrow{\mathcal{D}} 0$. La pregunta que uno podría hacerse a esta altura es a qué velocidad sucede esto. Es decir, si $a_n = n^k$ con k real, es una sucesión numérica que tiende a $+\infty$, que pasará si hacemos el producto: (recordar el Teorema de Slutsky)

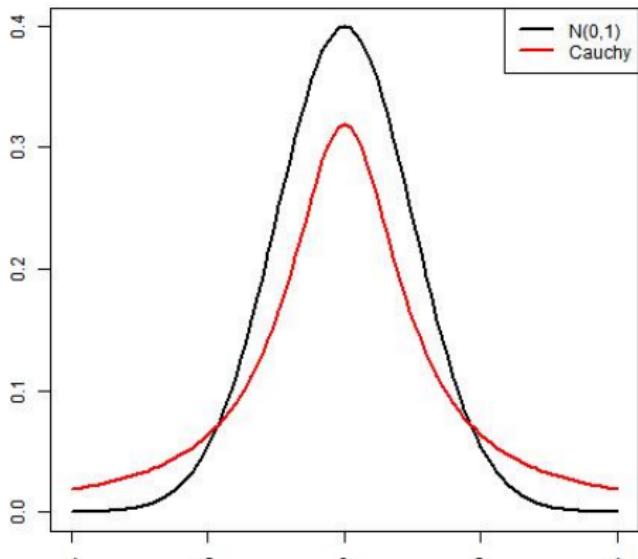
$$(\bar{X}_n - E(X_1)) a_n$$

¿Cómo tendrá que ser k para que el límite resultante dé algo que no sea ni cero ni infinito (y ojalá sea una variable aleatoria no constante)?

Y previamente, ¿habrá una tal velocidad? El TCL nos dice que si tomamos $a_n = \sqrt{n}$ entonces, $\sqrt{n} (\bar{X}_n - E(X_1)) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$

5. Sin el supuesto de que las X_i tengan **varianza finita**, el TCL falla, en general. El clásico contraejemplo es tomar las X_i con distribución Cauchy de parámetro $a > 0$, cuya densidad es

$$f_{X_i}(t) = \frac{a}{\pi(a^2 + t^2)}$$



En el gráfico vemos la densidad Cauchy con $a = 1$ junto con la densidad normal estándar. Cuando $a = 1$ se la denomina Cauchy estándar o t_1 , t de Student con un grado de libertad.

El promedio de v.a.i.i.d. Cauchy tendrá distribución Cauchy, también con parámetro a , y no puede converger a la normal. Pero esta distribución no tiene esperanza (ni varianza) finita. En contraste con la ley de los grandes números, incluso la condición $X_i \in \mathcal{L}^1$ no es suficiente, y tampoco podemos reemplazar el supuesto de independencia por independencia de a pares o incluso falta de correlación de a pares. Hay versiones más generales, entre las cuales la versión de Lindeberg y la versión de Lyapunov son las más importantes. También vale una versión del TCL para vectores aleatorios que se utiliza mucho en estadística.

Notación

El TCL se anota de diversas formas. Una posibilidad suele ser

$$S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{a} \mathcal{N}(0, 1),$$

donde el símbolo \xrightarrow{a} se lee “cuando n tiende a infinito su distribución es aproximadamente”

Otra notación es

$$\sum_{i=1}^n X_i \xrightarrow{a} \mathcal{N}(n\mu, n\sigma^2) \quad \text{cuando } n \text{ es grande,}$$

o sus versiones usando el promedio en vez de la suma.

Teorema Central del Límite para vectores aleatorios

Para dar una versión del TCL para vectores, debemos definir convergencia en distribución para vectores. Hay dos maneras de definirla, una es generalizar la convergencia de las funciones de distribución en los puntos de continuidad de la función de distribución del límite. La otra es tomar la equivalencia probada en el Teorema 7.23 y usarla como definición. Esta forma da una definición más operativa. Para ver la equivalencia entre ambas, puede consultarse el Durrett, R (2019) *Probability: Theory and Examples*. Misma fuente para las versiones de Lindeberg y Lyapunov.

Definición 8.1

Sea $(\mathbf{X}_i)_{i \geq 1}$, $\mathbf{X}_i \in \mathbb{R}^d$ una sucesión de vectores aleatorios y $\mathbf{X} \in \mathbb{R}^d$ otro vector. Diremos que $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ si y sólo si $E[g(\mathbf{X}_n)] \rightarrow E[g(\mathbf{X})]$ para toda función $g : \mathbb{R}^d \rightarrow \mathbb{R}$ continua y acotada.

Teorema Central del Límite para vectores aleatorios

Teorema 8.2 (TCL multivariado)

Sea $(\mathbf{X}_i)_{i \geq 1}$, $\mathbf{X}_i \in \mathbb{R}^d$ una sucesión de vectores aleatorios independientes e

idénticamente distribuidos con $E(\mathbf{X}_1) = \boldsymbol{\mu}$, con $\boldsymbol{\mu} = \begin{pmatrix} \mu(1) \\ \vdots \\ \mu(d) \end{pmatrix}$ y

covarianzas finitas que acomodamos en la matriz $\Sigma \in \mathbb{R}^{d \times d}$:

$\Sigma_{jk} := \text{cov}(\mathbf{X}_1(j), \mathbf{X}_1(k)) = E([\mathbf{X}_1(j) - \mu(j)][\mathbf{X}_1(k) - \mu(k)])$ Sea

$\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n$, entonces

$$\frac{\mathbf{S}_n - n\boldsymbol{\mu}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathbf{Y}$$

donde $\mathbf{Y} \sim \mathcal{N}_d(\mathbf{0}, \Sigma)$ es decir, tiene distribución normal multivariada d -dimensional de parámetros $\mathbf{0}$ y Σ , como definimos en la Definición 5.12.

Aun cuando los datos originales no tengan una distribución que parezca normal, sus promedios tendrán una distribución que se puede aproximar bien por la distribución normal. Por eso la distribución normal se usa en la práctica en muchos escenarios, incluso se la utiliza para aproximar probabilidades cuando el n no es demasiado grande (a veces está mal usada).

Veamos ejemplos de aplicaciones en casos particulares

- ① $X_i \sim Be(p)$
- ② Aplicación de binomiales: tablero de Galton
- ③ Aplicación de binomiales: ejemplo con la ruleta
- ④ Errores de medición
- ⑤ $X_i \sim \mathcal{U}(-1, 1)$

TCL para binomiales

Un caso de especial interés suele ser la aproximación que da el TCL cuando las X_i tienen distribución Bernoulli. Entonces $S_n = \sum_{i=1}^n X_i \sim Bi(n, p)$, y

$$\frac{S_n - np}{\sqrt{np(1-p)}} = \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Lo cual también puede escribirse en términos de los promedios, tenemos $E(\bar{X}_n) = p$ y $V(\bar{X}_n) = V(X_1)/n = \frac{p(1-p)}{n}$

$$\frac{\sqrt{n} (\bar{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

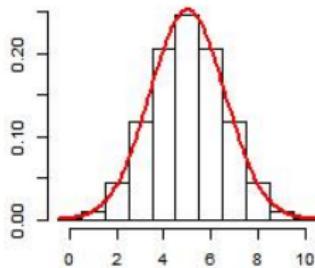
O, recordando propiedades,

$$\sqrt{n} (\bar{X}_n - p) \xrightarrow{\mathcal{D}} \mathcal{N}(0, p(1-p))$$

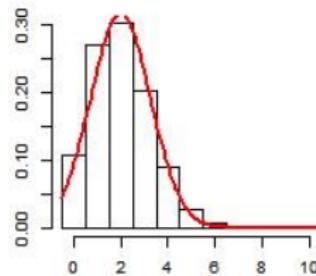
TCL para binomiales

Función de probabilidad puntual $B(n, p)$ para distintas combinaciones de n y p , y curva normal superpuesta

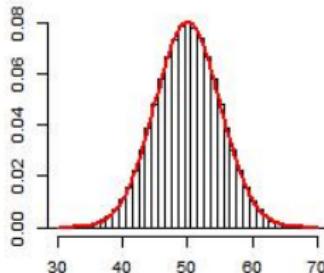
$n = 10, p = 0.5$



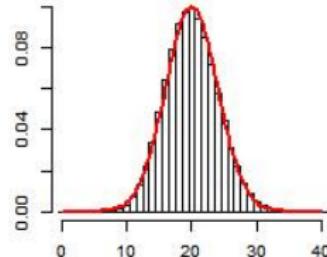
$n = 10, p = 0.2$



$n = 100, p = 0.5$



$n = 100, p = 0.2$



Tablero de Galton, o quincunx

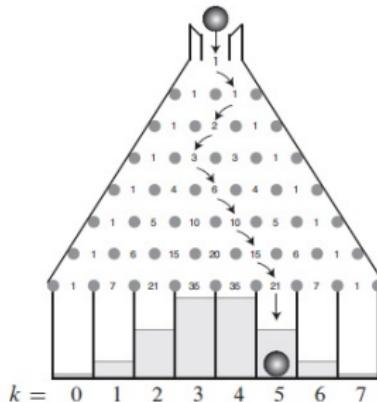


Figure 2.2. The Galton board for $n = 7$. The arrows mark a possible path for the ball. The numbers indicate the total number of paths leading to the respective position. Note that they form Pascal's triangle for the binomial coefficients. In the compartments at the bottom, the bar chart for $B_{7,1/2}$ is indicated in grey.

En acción <https://yihui.org/animation/example/quincunx/>
Video: <https://youtu.be/SZoDNfVFS7I>

Ruleta

Dudando del funcionamiento justo de una ruleta se lleva a cabo el siguiente experimento. Se repite 1000 veces tirar la bolilla de la ruleta, y se cuenta la cantidad de veces que se obtiene un resultado rojo. Se obtuvieron 430 resultados rojos. Con estos datos, ¿se mantiene la sospecha sobre la ruleta?

Sea X_i la variable indicadora de que la i ésima repetición arrojó un resultado rojo. Asumiendo que en el paño de la ruleta hay 38 números: los números del 1 al 36, mitad de los cuales son **rojos**, la otra mitad negros, y dos números **verdes**, el 0 y el 00. $X_i \sim Be(p)$ v.a.i.i.d. Si la ruleta es equilibrada $p = 18/38 = 0,4737$. Y en tal caso, podemos aproximar la probabilidad de obtener 430 resultados rojos, o menos aún.

Ruleta

Sea $S_n = X_1 + \dots + X_n$, $n = 1000$, $S_n \sim Bi(n, p)$.

Como $S_n \xrightarrow{d} \mathcal{N}(np, np(1 - p))$, si la ruleta fuera justa tendremos,

$$P(S_n \leq 430) = P\left(\frac{S_n - np}{\sqrt{np(1 - p)}} \leq \frac{430 - 1000 \times \frac{18}{38}}{\sqrt{1000 \times \frac{18}{38} \times \frac{20}{38}}}\right) \\ \approx \Phi(-2,77) = 0,00283$$

La probabilidad de observar el valor obtenido es muy baja. Por eso, concluimos que la ruleta no parece funcionar bien. Y la mandaríamos a arreglar. Esta aproximación al valor que queremos calcular es del orden de $\frac{1}{\sqrt{1000}} \approx 0,03$.

En este caso podemos calcular la probabilidad exacta, usando el R, con el comando `pbinom`,

```
> pbinom(430, size = 1000, p = 18/38)  
[1] 0.003070923
```

Errores de Medición

Suponga que X_1, \dots, X_n son mediciones repetidas, insesgadas e independientes de una cantidad, μ , y que la $V(X_i) = \sigma^2$. En las aplicaciones, la varianza suele ser conocida ya que depende del instrumento de medición. El promedio de las mediciones, \bar{X}_n , se usa como un estimador de μ . La Ley de los Grandes Números nos dice que \bar{X}_n converge a μ en probabilidad, de modo que podemos esperar que \bar{X}_n esté cerca de μ si n es grande. La desigualdad de Chebychev nos permite acotar la probabilidad de obtener un error de un tamaño determinado, pero el Teorema Central del Límite nos da una aproximación mucho más precisa del error real.

Deseamos encontrar $P(|\bar{X}_n - \mu| < c)$ para alguna constante c . Para usar el Teorema Central del Límite para aproximar esta probabilidad, primero estandarizamos, usando que $E(\bar{X}_n) = \mu$ y $V(\bar{X}_n) = \sigma^2/n$.

Errores de Medición

$$\begin{aligned} P(|\bar{X}_n - \mu| < c) &= P(-c < \bar{X}_n - \mu < c) \\ &= P\left(\frac{-c}{\sigma/\sqrt{n}} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < \frac{c}{\sigma/\sqrt{n}}\right) \\ &\approx \Phi\left(\frac{c\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{c\sqrt{n}}{\sigma}\right) \end{aligned}$$

Por ejemplo, supongamos que se toman 100 mediciones con $\sigma = 1$. La probabilidad de que el promedio discrepe de μ en menos de 0.2 es aproximadamente

$$P(|\bar{X}_n - \mu| < 0.2) \approx \Phi(0.2 \times 10) - \Phi(-0.2 \times 10) = 0.954$$

Este tipo de razonamiento también se suele invertir. Esto es, dado un nivel de precisión deseado en la discrepancia, c y una cota para la probabilidad δ , se puede despejar n para que

$$P(|\bar{X}_n - \mu| < c) \geq \delta$$

TCL para uniformes $\mathcal{U}[-1, 1]$

Sean $(X_i)_{i \geq 1}$, variables aleatorias independientes con distribución $\mathcal{U}[-1, 1]$.

- ① Es un ejercicio de la práctica de vectores aleatorios, probar que si $\overline{X}_2 = \frac{X_1 + X_2}{2}$, entonces

$$f_{\overline{X}_2}(u) = (u + 1) I_{(-1,0)}(u) + (1 - u) I_{(0,1)}(u).$$

- ② Sea $\overline{X}_3 = \frac{X_1 + X_2 + X_3}{3}$. Es un poco más trabajoso verificar que

$$f_{\overline{X}_3}(z) = \begin{cases} \frac{27}{16} (1 - |z|)^2 & \text{si } \frac{1}{3} < |z| < 1 \\ \frac{9}{8} - \frac{27}{8} z^2 & \text{si } |z| \leq \frac{1}{3} \\ 0 & \text{si } |z| \geq 1. \end{cases}$$

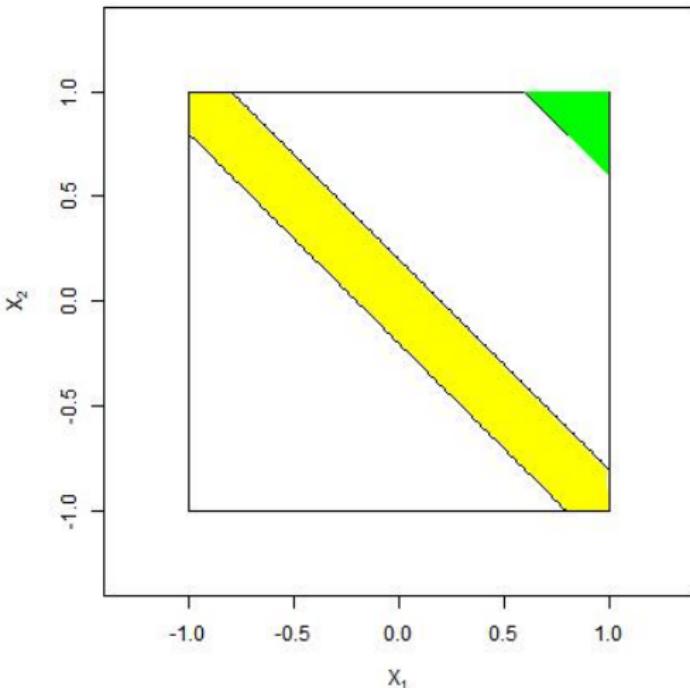
- ③ Observar que f_{X_1} no es continua en -1 y 1 , pero $f_{\overline{X}_2}$ es continua y $f_{\overline{X}_3}$ es derivable. Esto es una muestra de que la convolución mejora la densidad.

A la izquierda, el soporte del vector $(X_1, X_2) \sim \mathcal{U}([-1, 1] \times [-1, 1])$. Fijemos un $h = 0,1$, la idea es que después lo hagamos tender a cero.

Marcada en amarillo, la región a la que pertenece (X_1, X_2) para calcular
 $P(-h \leq \bar{X}_2 \leq h) = P(-X_1 - 2h \leq X_2 \leq -X_1 + 2h)$

En verde la región a la que pertenece (X_1, X_2) para calcular

$$P(1 - 2h \leq \bar{X}_2 \leq 1) = P(-X_1 + 2 - 4h < X_2 < 2 - X_1)$$



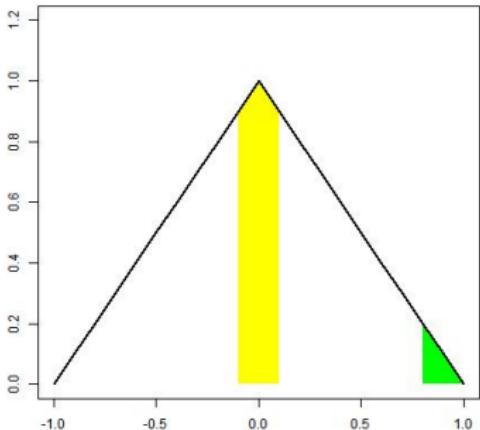
El gráfico de la densidad de \bar{X}_2 , con las probabilidades calculadas antes marcadas con su color.

Por el Lema 4.2, tenemos que si f_X es una función de densidad continua en x_0 , entonces

$$\lim_{h \rightarrow 0} \frac{P(X \in [x_0 - h, x_0 + h])}{2h} = f_X(x_0).$$

Para $x_0 = 0$, tenemos

$$\lim_{h \rightarrow 0} \frac{P(\bar{X}_2 \in [-h, h])}{2h} = \frac{2h(1-h/2)}{2h} = 1.$$



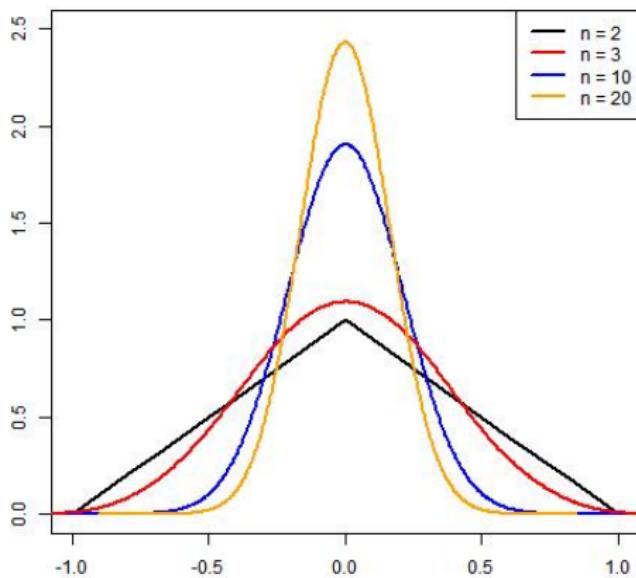
Para $x_0 = 1$, tenemos

$$\lim_{h \rightarrow 0} \frac{P(\bar{X}_2 \in [1-2h, 1])}{2h} = \frac{2h^2}{2h} = 0.$$

Densidad de \bar{X}_n para n vaid $\mathcal{U}[-1, 1]$

Vemos la LGN, las densidades se van concentrando alrededor de la esperanza, que es cero.

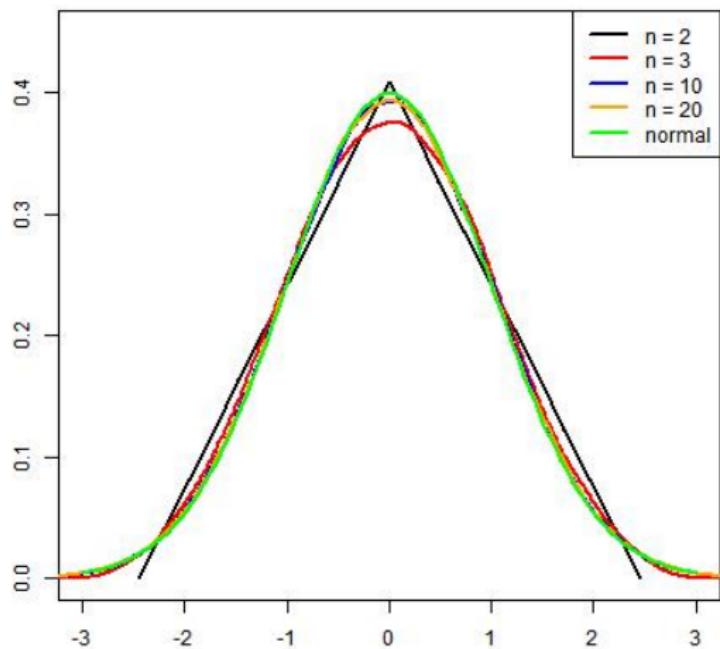
¿Por qué vemos la LGN, que es una convergencia en probabilidad, cuando graficamos densidades, que debería mostrar (si es posible visualizar) una convergencia en distribución? Por la jerarquía de convergencias, conv en probabilidad implica conv en distribución.



Densidad de \bar{X}_n para n vaid $\mathcal{U}[-1, 1]$

Estandarizadas según el TCL

Acá vemos que la función de densidad del promedio se va aproximando a la de la normal (se superponen tanto que no se distinguen)



Métrica para la Convergencia en Probabilidad

Ejercicio 8.A

(a) Mostrar que $d(X, Y) = E \left(\frac{|X - Y|}{1 + |X - Y|} \right)$ define una métrica en el conjunto de variables aleatorias, es decir,

- (i) $d(X, Y) \geq 0$
- (ii) $d(X, Y) = 0$ si y sólo si $X = Y$ c.s., o sea, $P(X = Y) = 1$.
- (iii) $d(X, Y) = d(Y, X)$
- (iv) $d(X, Z) \leq d(X, Y) + d(Y, Z)$.

(b) Mostrar que $d(X_n, X) \rightarrow 0$ cuando $n \rightarrow \infty$ si y sólo si $X_n \xrightarrow{P} X$.

Sugerencia: Recordar el Teorema 7.23, es decir, la caracterización de la convergencia en distribución vía esperanzas, para una de las implicaciones.

Para la otra, probar primero que $h(X_n) \xrightarrow{P} h(X)$ usando la Desigualdad de Markov, donde $h(t) = \frac{t}{t+1}$. Luego deducir la convergencia buscada usando el Teorema 7.17, es decir, el hecho de que la convergencia en probabilidad se preserva por funciones continuas.

Como consecuencia del Ejercicio 8.A, resulta que el conjunto de variables aleatorias es un espacio métrico, con la métrica d definida en él. Además, la convergencia en probabilidad es metrizable.

Recordemos el siguiente resultado de espacios métricos que relaciona la noción de límite con convergencia de sucesiones.

Teorema 8.A

Sea $g : E \rightarrow M$ donde E y M son espacios métricos. Las siguientes dos afirmaciones son equivalentes:

- $\lim_{x \rightarrow x_0} g(x) = L$
- Para toda sucesión $x_n \rightarrow x_0$ se tiene $\lim_{n \rightarrow \infty} g(x_n) = L$.

Cuando trabajamos con funciones características, definimos funciones $Y : \mathbb{R} \rightarrow \mathcal{A}$ con \mathcal{A} el espacio de variables aleatorias, dotado de la métrica definida en el Ejercicio 8.A. En particular, $\lim_{h \rightarrow 0} Y(h) = Y$ en probabilidad, si y sólo si $Y(a_n) \xrightarrow{P} Y$ para toda sucesión $(a_n)_{n \geq 1}$ que cumpla $a_n \rightarrow 0$ cuando $n \rightarrow \infty$. Usamos esta equivalencia en la demostración del Lema 8.2(5) y en la Proposición 8.1.

Función Característica

Definición 8.1

La función característica de la variable aleatoria X es una función $\varphi : \mathbb{R} \rightarrow \mathbb{C}$ dada por

$$\varphi_X(t) = E[\cos(tX)] + iE[\sin(tX)] = E[e^{itX}]$$

Propiedades de los números complejos

Recordemos algunas propiedades de los números complejos. Dado $z = a + bi \in \mathbb{C}$,

- La parte real, $\operatorname{Re}(z) = a$. La parte compleja, $\operatorname{Im}(z) = b$.
- El módulo de z es $|z| = \sqrt{a^2 + b^2}$.
- El argumento de z , θ es el ángulo, dado en radianes, en el plano euclíadiano entre el eje x positivo y el rayo desde el origen al punto $(a, b) \neq (0, 0)$, medido en el sentido antihorario.
- La fórmula de Euler relaciona estas cantidades,
$$z = |z| e^{i\theta} = |z| (\cos(\theta) + i \sin(\theta)).$$
- El conjugado de z , \bar{z} , se define por $\bar{z} = a - bi$.
- Vale $|z|^2 = z\bar{z}$.
- z es real si y sólo si $\bar{z} = z$.

Lema 8.1

$$\left| e^{ix} - \sum_{m=0}^n \frac{(ix)^m}{m!} \right| \leq \min \left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right)$$

El primer término de la derecha es el orden usual de magnitud que esperamos en un término de error. El segundo es mejor para $|x|$ grande, y nos ayudará a probar el TCL sin asumir el tercer momento finito.

Demostración.

Integrando por partes obtenemos

$$\int_0^x (x-s)^n e^{is} ds = \frac{x^{n+1}}{n+1} + \frac{i}{n+1} \int_0^x (x-s)^{n+1} e^{is} ds \quad (1)$$

Cuando $n = 0$, esto dice

$$\int_0^x e^{is} ds = x + i \int_0^x (x-s) e^{is} ds$$

Demostración.

El lado izquierdo es $\int_0^x e^{is} ds = (e^{ix} - 1) / i$. Hagamos la cuenta en \mathbb{R} para verificarlo, sea $h(s) = e^{is} = \cos(s) + i \sin(s)$,
 $h'(s) = -\sin(s) + i \cos(s) = i (\cos(s) - \frac{1}{i} \sin(s)) = i (\cos(s) + i \sin(s)) = ie^{is}$.

Luego reacomodando tenemos

$$e^{ix} = 1 + ix + i^2 \int_0^x (x-s)e^{is} ds$$

Usando el resultado para $n = 1$ nos da

$$e^{ix} = 1 + ix + \frac{i^2 x^2}{2} + \frac{i^3}{2} \int_0^x (x-s)^2 e^{is} ds$$

e iterando obtenemos

$$e^{ix} - \sum_{m=0}^n \frac{(ix)^m}{m!} = \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} ds \quad (2)$$

Demostración.

Para probar el resultado, sólo resta acotar el “término del error” del lado derecho. Como $|e^{is}| \leq 1$ para todo s

$$\left| \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} ds \right| \leq |x|^{n+1} / (n+1)! \quad (3)$$

La última cota es buena cuando x es chico. La que sigue funciona bien para x grande. Reescribimos la ecuación (1) con $n-1$ en el lugar de n

$$\frac{i}{n} \int_0^x (x-s)^n e^{is} ds = -\frac{x^n}{n} + \int_0^x (x-s)^{n-1} e^{is} ds$$

Observemos que $x^n/n = \int_0^x (x-s)^{n-1} ds$, entonces

$$\frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} ds = \frac{i^n}{(n-1)!} \int_0^x (x-s)^{n-1} (e^{is} - 1) ds$$

Demostración.

Como $|e^{ix} - 1| \leq 2$, sigue que

$$\left| \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} ds \right| \leq \left| \frac{2}{(n-1)!} \int_0^x (x-s)^{n-1} ds \right| \leq 2|x|^n/n! \quad (4)$$

Combinando (2), (3), y (4) obtenemos el resultado buscado. □

Algunas desigualdades importantes que se deducen del Lema 8.1

① $|e^{i\alpha} - 1| \leq |\alpha|$

② $|e^{i\alpha} - e^{i\beta}| \leq |\alpha - \beta|$

Dem: $|e^{i\alpha} - e^{i\beta}| = |e^{i\beta} (e^{i(\alpha-\beta)} - 1)| = |e^{i\beta}| |e^{i(\alpha-\beta)} - 1| \leq |\alpha - \beta|$,
por (1).

Lema 8.2

Sea X una variable aleatoria y $\varphi_X(t)$ su función característica.

- ① $|\varphi_X(t)| \leq 1$ para todo $t \in \mathbb{R}$. Dem: $|\varphi_X(t)| \leq E[|e^{itX}|] = 1$.
- ② $\varphi_X(0) = 1$.
- ③ $\overline{\varphi_X(t)} = \varphi_X(-t)$
- ④ $\overline{\varphi_X}$ es la función característica de $-X$: $\overline{\varphi_X} = \varphi_{-X}$
- ⑤ φ_X es uniformemente continua. Dem: Tenemos que
 $|\varphi_X(t+h) - \varphi_X(t)| = |E(e^{i(t+h)X} - e^{itX})| \leq E|e^{itX}(e^{ihX} - 1)| \leq E(Y(h))$ con $Y(h) = |e^{ihX} - 1|$. Queremos probar que
 $E(Y(h)) \xrightarrow{h \rightarrow 0} 0 \Leftrightarrow E(Y(a_n)) \xrightarrow{n \rightarrow \infty} 0$ para toda sucesión $(a_n)_{n \geq 1}$ tal que $a_n \xrightarrow{n \rightarrow \infty} 0$. Pero $|Y(h)| \leq 2$ e $Y(h) \xrightarrow{P} 0$ cuando $h \rightarrow 0$. Es decir, para toda sucesión $(a_n)_{n \geq 1}$ tal que $a_n \xrightarrow{n \rightarrow \infty} 0$, la sucesión de v.a. dada por $(Y(a_n))_{n \geq 1}$ converge en probabilidad a 0 cuando $n \rightarrow \infty$. Luego, por el Teorema de la convergencia acotada, resulta $E(Y(a_n)) \xrightarrow{n \rightarrow \infty} 0$.

Lema 8.2 (cont.)

Sea X una variable aleatoria y $\varphi_X(t)$ su función característica.

6) Si $Y = aX + b$, entonces $\varphi_Y(t) = e^{itb}\varphi_X(at)$ Dem:

$$\begin{aligned}\varphi_Y(t) &= \varphi_{aX+b}(t) = E(e^{it(aX+b)}) = E(e^{itaX}e^{itb}) = \\ &e^{itb}E(e^{itaX}) = e^{itb}\varphi_X(at), \text{ para todo } t \in \mathbb{R}.\end{aligned}$$

Ejemplo 8.1 (Característica de una $\mathcal{N}(0,1)$)

Si $Z \sim N(0, 1)$, $f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ y

$$\varphi_Z(t) = E[e^{itZ}] = \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Como

$$\begin{aligned} -\frac{x^2}{2} + itx &= -\left[\left(\frac{x}{\sqrt{2}}\right)^2 - 2\left(\frac{x}{\sqrt{2}}\right)\left(\frac{\sqrt{2}it}{2}\right) + \frac{2i^2t^2}{4}\right] + \frac{1}{2}i^2t^2 = \\ &= -\left[\frac{x}{\sqrt{2}} - \frac{1}{\sqrt{2}}it\right]^2 - \frac{t^2}{2}, \text{ completando cuadrados tenemos} \end{aligned}$$

$$\begin{aligned} \varphi_Z(t) &= \int_{-\infty}^{\infty} e^{-t^2/2} e^{-\left[\frac{x}{\sqrt{2}} - \frac{1}{\sqrt{2}}it\right]^2} \frac{1}{\sqrt{2\pi}} dx \\ &= e^{-t^2/2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\left[\frac{x}{\sqrt{2}} - \frac{1}{\sqrt{2}}it\right]^2} dx}_{=1 \text{ "densidad" de una } \mathcal{N}(it, 1)} = e^{-t^2/2} \end{aligned}$$

Ejemplo 8.1, cont. (Característica de una $\mathcal{N}(0, 1)$)

La forma correcta de probar esta igualdad, que la integral de una "densidad" de una $\mathcal{N}(it, 1)$ da 1 es vía el Teorema de Cauchy, que verán en Análisis Complejo, y luego tomando límites. O bien,

$$\varphi_Z(t) = \int_{-\infty}^{\infty} \cos(tx) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx + i \underbrace{\int_{-\infty}^{\infty} \sin(tx) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx}_{=0 \text{ por ser impar}}$$

Observar que $\varphi'_Z(t) =$

$$\int_{-\infty}^{\infty} -x \sin(tx) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \stackrel{\text{partes}}{=} \int_{-\infty}^{\infty} -t \cos(tx) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = -t \varphi_Z(t)$$

Luego $\frac{\varphi'_Z(t)}{\varphi_Z(t)} = -t$, integrando se obtiene

$\ln(\varphi_Z(t)) = -\frac{t^2}{2} + c \iff \varphi_Z(t) = \exp\left(-\frac{t^2}{2} + c\right)$. Como $\varphi_Z(0) = 1$, resulta $c = 0$. Otra demostración puede encontrarse en el apunte de Víctor J. Yohai, http://cms.dm.uba.ar/academico/materias/1ercuat2019/probabilidades_y_estadistica_M/apunte2.pdf

Ejemplo 8.2

Si $P(X = 1) = P(X = -1) = 1/2$, entonces

$$\varphi_X(t) = E\left(e^{itX}\right) = (e^{it} + e^{-it}) \frac{1}{2} = \cos(t)$$

Como X es simétrica alrededor de cero, es decir, X y $-X$ tienen la misma distribución, entonces resulta que $\varphi_X(t)$ es real. Lo mismo sucede con la función característica de la $\mathcal{N}(0, 1)$ que calculamos en el ejemplo anterior.

Ejercicio 8.1

- Probar que si $X \sim \text{Poisson}(\lambda)$ entonces $\varphi_X(t) = \exp(\lambda(e^{it} - 1))$.
- Probar que si $X \sim \text{Exp}(1)$ entonces $\varphi_X(t) = \frac{1}{1-it}$ y hallar φ_Y para $Y \sim \text{Exp}(\lambda)$
- A partir de φ_Z de $Z \sim \mathcal{N}(0, 1)$ calcular φ_X para $X \sim \mathcal{N}(\mu, \sigma^2)$.

El próximo teorema da la razón por la cual se introducen las funciones características en probabilidades.

Teorema 8.3

Sean X e Y dos variables aleatorias independientes. Entonces

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) \text{ para todo } t \in \mathbb{R}.$$

Demostración.

Observando que $\exp(itX)$, $\exp(itY)$ son variables aleatorias independientes,

$$\begin{aligned}\varphi_{X+Y}(t) &= E(\exp(it(X+Y))) \\ &= E(\exp(itX)\exp(itY)) \\ &= E(\exp(itX))E(\exp(itY)) = \varphi_X(t)\varphi_Y(t).\end{aligned}$$

El Teorema 8.3 nos dice que si $X \sim \mathcal{N}(0, \sigma_X^2)$ e $Y \sim \mathcal{N}(0, \sigma_Y^2)$ son independientes, entonces $\varphi_{X+Y}(t) = e^{-t^2(\sigma_X^2 + \sigma_Y^2)/2}$. ¿Nos dice esto que $X + Y$ tiene distribución normal? (Cosa que ya sabemos, por otro lado.) No, nos faltan un par de resultados todavía.

Relación entre función característica y momentos

Recordemos que dada una variable aleatoria X definíamos el momento k -ésimo de X como la $E(X^k)$. Nos había quedado un resultado de contención del capítulo anterior sin probar. Retomémoslo.

Lema 7.7 (debíamos la prueba)

Si $1 \leq p < q$, entonces $\mathcal{L}^q \subset \mathcal{L}^p$, es decir que si $E|X|^q < \infty$ entonces $E|X|^p < \infty$ para todo $p < q$.

Lema 7.7 (debíamos la prueba)

Si $1 \leq p < q$, entonces $\mathcal{L}^q \subset \mathcal{L}^p$, es decir que si $E|X|^q < \infty$ entonces $E|X|^p < \infty$ para todo $p < q$.

Demostración.

Sea $p < q$. Se tiene

$$\begin{aligned}|X|^p &= I_{\{|X| \leq 1\}} |X|^p + I_{\{|X| > 1\}} |X|^p \\&\leq I_{\{|X| \leq 1\}} + I_{\{|X| > 1\}} |X|^q \\&\leq I_{\{|X| \leq 1\}} + |X|^q\end{aligned}$$

Tomando esperanza en ambos miembros resulta

$$E(|X|^p) \leq P(|X| \leq 1) + E(|X|^q) < \infty,$$

y esto prueba el resultado. □

Relación entre función característica y momentos

Esta herramienta será usada en el resultado que sigue (omitimos la prueba). Generaliza al teorema de la convergencia acotada.

Teorema 8.4 (Teorema de la convergencia dominada)

Sea $(X_n)_{n \geq 1}$ una sucesión de variables aleatorias tales que $X_n \xrightarrow{P} X$. Si existe $Y > 0$ con $E[Y] < \infty$ tal que $|X_n| \leq Y$ para todo n , entonces

$$E[X_n] \rightarrow E[X]$$

Para una prueba puede verse el apunte de Víctor Yohai.

Relación entre la función característica y momentos

Proposición 8.1

Si $E[|X|^k] < \infty$, entonces φ_X tiene hasta k derivadas continuas acotadas: $\varphi_X \in \mathcal{C}_\infty^k$. Mas aún, vale que

$$\varphi_X^{(m)}(t) = E[e^{itX}(iX)^m], \forall m \leq k. \quad (5)$$

En particular,

$$\varphi_X^{(m)}(0) = i^m E[X^m], \forall m \leq k. \quad (6)$$

Demostración.

$X \in \mathcal{L}^k$, y por lo tanto en \mathcal{L}^1 . Comencemos mirando la primer derivada en detalle. Tenemos

$$\varphi'_X(t) = \lim_{h \rightarrow 0} \frac{\varphi_X(t+h) - \varphi_X(t)}{h} = \lim_{h \rightarrow 0} \frac{1}{h} E \left[e^{i(t+h)X} - e^{itX} \right]$$

Las variables aleatorias $Z_h = (1/h) [e^{i(t+h)X} - e^{itX}]$ convergen a $Z = iXe^{itX}$ cuando $h \rightarrow 0$. Y, más aún, como $|e^{ity} - e^{itz}| \leq |y - z|t$ como consecuencia del Lema 8.1, tenemos

$$|Z_h| = \frac{|e^{i(t+h)X} - e^{itX}|}{|h|} \leq \frac{|h||X|}{|h|} = |X|$$

por lo que las Z_h están dominadas por $|X| \in \mathcal{L}^1$. Luego, por el Teorema de la convergencia dominada tenemos

$$\lim_{h \rightarrow 0} E[Z_h] = E \left[\lim_{h \rightarrow 0} Z_h \right] = E \left[(iX)e^{itX} \right]$$

Demostración.

Entonces, probamos $\varphi'_X(t) = E [(iX)e^{itX}]$ Repitiendo este argumento de manera inductiva se obtiene el resultado.

$P(m)$: Si $E(|X|^m) < \infty$ entonces $\varphi_X^{(m)}(t) = E [(iX)^m e^{itX}]$ Vamos a probar $P(m+1)$, es decir, sabiendo que $E(|X|^{m+1}) < \infty$ calculamos

$$\begin{aligned}\varphi_X^{(m+1)}(t) &= \lim_{h \rightarrow 0} \frac{\varphi_X^{(m)}(t+h) - \varphi_X^{(m)}(t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{1}{h} E \left[i^m X^m \left(e^{i(t+h)X} - e^{itX} \right) \right]\end{aligned}$$

Observar que hemos usado la hipótesis inductiva en la segunda igualdad. Como antes, llamando $Z_h = (1/h) [e^{i(t+h)X} - e^{itX}]$, tenemos

$$\varphi_X^{(m+1)}(t) = \lim_{h \rightarrow 0} E [i^m X^m Z_h]$$



Demostración.

(copiamos)

$$\varphi_X^{(m+1)}(t) = \lim_{h \rightarrow 0} E[i^m X^m Z_h]$$

Vimos que $|Z_h| \leq |X| \in \mathcal{L}^1$, entonces

$|i^m X^m Z_h| \leq |X|^m |Z_h| \leq |X|^{m+1} \in \mathcal{L}^1$ por hipótesis. Además, Z_h converge a $Z = iX e^{itX}$ cuando $h \rightarrow 0$. Luego, nuevamente por el Teorema de la Convergencia Mayorada, tenemos

$$\begin{aligned}\varphi_X^{(m+1)}(t) &= \lim_{h \rightarrow 0} E[i^m X^m Z_h] = E\left[\lim_{h \rightarrow 0} i^m X^m Z_h\right] \\ &= E\left[i^m X^m \lim_{h \rightarrow 0} Z_h\right] = E\left[(iX)^{m+1} e^{itX}\right]\end{aligned}$$

lo cual prueba $P(m+1)$. Y por lo tanto (5). (6) sale evaluando en 0. □

Ejemplo 8.3

Calculemos los primeros cuatro momentos de una distribución $\mathcal{N}(\mu, \sigma^2)$. Para eso, calculamos primero los de $Z \sim \mathcal{N}(0, 1)$, usando la Proposición 8.1.

$$\begin{aligned}\varphi_Z(t) &= e^{-t^2/2} & \varphi'_Z(t) &= e^{-t^2/2}(-t) \\ \varphi_Z^{(2)}(t) &= e^{-\frac{1}{2}t^2}(t^2 - 1) & \varphi_Z^{(3)}(t) &= e^{-\frac{1}{2}t^2}(-t^3 + 3t) \\ \varphi_Z^{(4)}(t) &= e^{-t^2/2}(t^4 - 6t^2 + 3)\end{aligned}$$

Evaluando en cero tenemos: $\varphi_Z(0) = 1$, $\varphi'_Z(0) = 0$, $\varphi_Z^{(2)}(0) = -1$, $\varphi_Z^{(3)}(0) = 0$, y $\varphi_Z^{(4)}(0) = 3$. Luego, de $\varphi_Z^{(m)}(0) = E[(iZ)^m]$ resulta, usando que $\frac{1}{i} = -i$,

$$E[Z^m] = (-i)^m \varphi_Z^{(m)}(0),$$

por lo que $E[Z] = 0$, $E[Z^2] = 1$, $E[Z^3] = 0$, $E[Z^4] = 3$.

Usando que $X = \mu + \sigma Z$ pueden obtenerse los momentos de X ,

$$\begin{aligned}E[X] &= \mu, & E[X^2] &= \mu^2 + \sigma^2, & E[X^3] &= \mu^3 + 3\mu\sigma^2, \\ E[X^4] &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4\end{aligned}$$

Desarrollo de Taylor de la función característica

Proposición 8.2

Taylor para características Si $E[|X|^k] < \infty$, entonces φ_X tiene hasta k derivadas continuas acotadas: $\varphi_X \in C_\infty^k$. Y vale que

$$\varphi_X(t) = \sum_{m=0}^k \frac{i^m}{m!} E[X^m] t^m + R(t) \quad (7)$$

con $R(t) = o(|t|^k)$, es decir, R satisface que $\lim_{t \rightarrow 0} \frac{R(t)}{|t|^k} = 0$.

Demostración.

A partir de la Proposición 8.1, la ecuación (6) nos da los coeficientes del polinomio de Taylor de φ_X desarrollado alrededor de cero:

$$\varphi_X^{(m)}(0) = i^m E[X^m], \forall m \leq k.$$

Demostración.

Queremos acotar

$$\begin{aligned}|R(t)| &= \left| \varphi_X(t) - \sum_{m=0}^k \frac{i^m}{m!} E[X^m] t^m \right| = \left| Ee^{itX} - \sum_{m=0}^k E\left(\frac{(itX)^m}{m!}\right) \right| \\&= \left| E \left[e^{itX} - \sum_{m=0}^k \frac{(itX)^m}{m!} \right] \right| \leq E \left| e^{itX} - \sum_{m=0}^k \frac{(itX)^m}{m!} \right|\end{aligned}$$

Por el Lema 8.1, tenemos que

$$\left| e^{i\textcolor{orange}{x}} - \sum_{m=0}^k \frac{(i\textcolor{orange}{x})^m}{m!} \right| \leq \min \left(\frac{|\textcolor{orange}{x}|^{k+1}}{(k+1)!}, \frac{2|\textcolor{orange}{x}|^k}{k!} \right)$$

Usando el Lema 8.1 para $\textcolor{orange}{x} = tX$, tenemos

$$|R(t)| \leq E \min \left(\frac{|tX|^{k+1}}{(k+1)!}, \frac{2|tX|^k}{k!} \right) = |t|^k E \left[\min \left(\frac{|t||X|^{k+1}}{(k+1)!}, \frac{2|X|^k}{k!} \right) \right]$$

Demostración.

(copiamos)

$$|R(t)| \leq |t|^k E \left[\min \left(\frac{|t||X|^{k+1}}{(k+1)!}, \frac{2|X|^k}{k!} \right) \right]$$

La variable **entre corchetes** está acotada por $\frac{2|X|^k}{k!}$ que tiene esperanza finita, y además, está acotada por $\frac{|tX|^{k+1}}{(k+1)!}$ que tiende a cero cuando $t \rightarrow 0$. Luego, por el Teorema de la Convergencia dominada, resulta que $\lim_{t \rightarrow 0} \frac{|R(t)|}{|t|^k} = 0$.

□

Observación 8.1

El objetivo de escribir la cota del Lema 8.1 que involucra al mínimo de dos términos en vez de sólo el primero que resultaría de una aplicación ingenua del desarrollo de Taylor es que obtenemos la conclusión de la Proposición 8.2 bajo el supuesto de que $E[|X|^k] < \infty$ sin necesidad de asumir $E[|X|^{k+1}]$. Esto es crucial al probar el TCL o la versión débil de la LGN que se pueden hacer usando funciones características.

Teorema 8.5 (Teorema de inversión)

Sea X una v.a. Para todo $a < b$ puntos de continuidad de F_X vale que

$$F_X(b) - F_X(a) = \frac{1}{2\pi} \lim_{R \rightarrow \infty} \int_{-R}^R \frac{e^{-iat} - e^{-ibt}}{it} \varphi_X(t) dt$$

Demostración.

(caso continuo) Sea f_X la densidad, consideremos

$$I(R) = \int_{-R}^R \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt = \int_{-R}^R \frac{e^{-ita} - e^{-itb}}{it} \left[\int_{-\infty}^{\infty} e^{itx} f_X(x) dx \right] dt$$

Para poder cambiar el orden de integración por Fubini, basta que la integral sea finita. Como

$|e^{it(x-a)} - e^{it(x-b)}| \leq |t(x-a) - t(x-b)| = |t(b-a)|$ resulta
 $\left| \frac{e^{it(x-a)} - e^{it(x-b)}}{it} \right| f_X(x) \leq (b-a)f_X(x)$ que es integrable en $[-R, R] \times (-\infty, \infty)$.



Demostración.

Entonces $I(R) = \int_{-\infty}^{\infty} \left[\int_{-R}^R \frac{e^{it(x-a)} - e^{it(x-b)}}{it} dt \right] f_X(x) dx$. Consideremos ahora

$$g_R(x) = \int_{-R}^R \frac{e^{it(x-a)} - e^{it(x-b)}}{it} dt = 2 \int_0^R \frac{\sin(t(x-a))}{t} - \frac{\sin(t(x-b))}{t} dt$$

Entonces tenemos

$$I(R) = E[g_R(X)]$$

Queremos hallar el $\lim_{R \rightarrow \infty} I(R)$. Vamos a aplicar el teorema de la convergencia acotada (Prop 7.7) al término de la derecha. Para ello, veremos que

- $g_R(x) \rightarrow g_\infty(x)$ para todo x , y además
- $|g_R(x)| \leq M$, para todo R , para todo x .

Luego, tendremos que

$$\lim_{R \rightarrow +\infty} E[g_R(X)] = E[g_\infty(X)] .$$

Demostración.

Usaremos que $\int_0^R \frac{\sin(\theta t)}{t} dt = \operatorname{sgn}(\theta) S(R|\theta|)$, para $R > 0$, donde

$$\operatorname{sgn}(\theta) = \begin{cases} 1 & \text{si } \theta > 0 \\ -1 & \text{si } \theta < 0 \\ 0 & \text{sino} \end{cases}$$

y $S(R) = \int_0^R \frac{\sin(t)}{t} dt$. Vale que $\lim_{R \rightarrow +\infty} S(R) = \frac{\pi}{2}$. Para una prueba, puede verse Billingsley, P. (1995) *Probability and Measure*. John Wiley & Sons, Inc. 3rd edition. Ejemplo 18.4.

Además, $S(R)$ está acotada. Entonces,

$$\begin{aligned} g_R(x) &= 2 \int_0^R \frac{\sin(t(x-a))}{t} - \frac{\sin(t(x-b))}{t} dt \\ &= 2 \operatorname{sgn}(x-a) S(R|x-a|) - 2 \operatorname{sgn}(x-b) S(R|x-b|) \end{aligned}$$



Demostración.

(copiamos) $g_R(x) = 2 \operatorname{sgn}(x - a) S(R|x - a|) - 2 \operatorname{sgn}(x - b) S(R|x - b|)$

Concluimos que existe M tal que $|g_R(z)| \leq M$ y cuando $R \rightarrow +\infty$,

$$g_R(x) \rightarrow g_\infty(x) = \begin{cases} 0 & \text{si } x < a , \\ \pi & \text{si } x = a , \\ 2\pi & \text{si } a < x < b , \\ \pi & \text{si } x = b , \\ 0 & \text{si } x > b . \end{cases}$$

En otras palabras,

$$g_R(X) \rightarrow \pi I_{\{X=a\}} + 2\pi I_{\{a < X < b\}} + \pi I_{\{X=b\}}$$

Finalmente, tenemos que

$$\lim_{R \rightarrow \infty} E[g_R(X)] = E[g_\infty(X)] = E[g_\infty(X)] = 2\pi\{F_X(b) - F_X(a)\}$$

siendo a y b puntos de continuidad de F_X .

La consecuencia más importante del Teorema de Inversión es que la distribución de una v.a. queda únicamente determinada por su función característica.

Teorema 8.6

Si $\varphi_X(t) = \varphi_Y(t)$ para todo $t \in \mathbb{R}$ entonces $F_X = F_Y$.

Demostración.

Si a y b son puntos de continuidad simultáneamente de F_X y de F_Y , tenemos

$$F_X(b) - F_X(a) = F_Y(b) - F_Y(a).$$

Tomando $\lim_{a \rightarrow -\infty}$ tenemos:

$$F_X(b) = F_Y(b),$$

para todo b punto de continuidad de F_X y de F_Y . Como las funciones de distribución tienen a lo sumo numerables puntos de discontinuidad, resulta que $F_X = F_Y$. □

Teorema 8.7 (Teorema de inversión de Fourier)

Si $\int_{-\infty}^{\infty} |\varphi_X(t)| dt < \infty$ entonces X es absolutamente continua con densidad

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t) dt$$

Demostración.

$$\begin{aligned} F_X(b) - F_X(a) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_a^b e^{-itx} dx \right] \varphi_X(t) dt \\ &= \int_a^b \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t) dt \right] dx \end{aligned}$$

Usamos Fubini porque el valor absoluto del integrando,
 $\left| \frac{1}{2\pi} e^{-itx} \varphi_X(t) \right| \leq \frac{1}{2\pi} |\varphi_X(t)|$ es integrable por hipótesis. □

Demostración.

Comprobemos la “primitiva” compleja. Hagamos la cuenta en \mathbb{R} , sea

$$h(x) = \frac{-e^{-itx}}{it} = -\left[\frac{\cos(-tx)+i\sin(-tx)}{it}\right],$$

$$\begin{aligned} h'(x) &= -\left[\frac{t\sin(-tx)+i(-t)\cos(-tx)}{it}\right] = -\left[\frac{1}{i}\sin(-tx)-\cos(-tx)\right] = \\ &\cos(-tx)+i\sin(-tx)=e^{-itx}. \end{aligned}$$

Luego

$$\int_a^b e^{-itx} dx = \left[\frac{-e^{-itx}}{it} \right] \Big|_a^b = \frac{e^{-ita}-e^{-itb}}{it}$$



Ejemplo 8.4 (Distribución Normal)

Si $Z \sim N(0, 1)$, con $\varphi_Z(t) = e^{-t^2/2}$, aplicando el resultado anterior tenemos (completando cuadrados)

$$\begin{aligned}f_Z(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} e^{-t^2/2} dt \\&= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x+it)^2/2} dt \\&= \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.\end{aligned}$$

Teorema 8.8 (Teorema de Continuidad)

Sea $(X_n)_{n \geq 1}$ una sucesión de variables aleatorias con funciones de distribución acumulada F_{X_n} y funciones características φ_{X_n} , respectivamente. Entonces

- i. $X_n \xrightarrow{\mathcal{D}} X \Rightarrow \varphi_{X_n}(t) \rightarrow \varphi_X(t), \forall t \in \mathbb{R};.$
- ii. Si $\varphi_{X_n}(t) \rightarrow \varphi(t), \forall t \in \mathbb{R}$ y $\varphi : \mathbb{R} \rightarrow \mathbb{C}$ es continua en cero. Entonces, la sucesión $(X_n)_{n \geq 1}$ es acotada en probabilidad y φ , el límite, es la función característica de una distribución F . Si $Y \sim F$ entonces $X_n \xrightarrow{\mathcal{D}} Y$.

Demostración.

- i Para todo $t \in \mathbb{R}$ las funciones $a_t(x) = \cos(tx)$ y $b_t(x) = \sin(tx)$ son ambas continuas y acotadas. El resultado sigue del Teorema 7.23 que nos dice que la convergencia en distribución es equivalente a que $E[g(X_n)] \rightarrow E[g(X)]$ para toda g continua y acotada.
- ii (Sin demostración). Puede verse Durrett, R. (2019) *Probability Theory and Examples*, Teorema 3.3.17 o Billingsley, P. (1995) *Probability and Measure*, Teorema 26.3 y sus corolarios.

Volvamos al

Teorema 8.1 (Teorema Central del Límite)

Sean $(X_i)_{i \geq 1}$ v.a.i.i.d. con $E[X_1] = \mu$ y $V(X_1) = \sigma^2 < +\infty$, entonces tenemos que

$$\frac{\sum_{i=1}^n X_i - E(\sum_{i=1}^n X_i)}{\sqrt{V(\sum_{i=1}^n X_i)}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{\mathcal{D}} Z \quad Z \sim \mathcal{N}(0, 1),$$

Ejercicio 8.2

Probar el TCL usando funciones características y asumiendo el siguiente resultado como válido (que extiende a los complejos un resultado conocido para sucesiones reales).

Teorema 8.9

Si $(c_n)_{n \geq 1}$ es una sucesión numérica tal que $c_n \rightarrow c$ con $c_n, c \in \mathbb{C}$ entonces $(1 + c_n/n)^n \rightarrow e^c$.

Teorema 8.10 (LGN débil, con primer momento finito)

Sean $(X_i)_{i \geq 1}$ v.a.i.i.d, si $E[|X_1|] < +\infty$, entonces

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} E[X_1].$$

Ejercicio 8.3

Probar la LGN, versión débil para primer momento finito usando funciones características.

Ejercicio 8.4

Sean $(X_n)_{n \geq 1}$ variables aleatorias i.i.d. con $E[X_i] = \mu$ y $V(X_i) = \sigma^2$.

- ① Probar que la varianza muestral $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ converge casi seguramente a σ^2 . Sugerencia: escribir a S^2 como una resta de variables aleatorias, desarrollando el cuadrado.
- ② Probar que

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{\mathcal{D}} Z \quad \text{con } Z \sim \mathcal{N}(0, 1).$$

Sugerencia: Recordar Slutsky.

Ejercicio 8.5

(Propagación de errores para variables transformadas o método delta)

Sean $(a_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}$ una sucesión que satisface $\lim_{n \rightarrow +\infty} a_n = +\infty$ y $(X_n)_{n \in \mathbb{N}}$ una sucesión de variables aleatorias tal que $a_n(X_n - \mu) \xrightarrow{\mathcal{D}} X$ para un determinado parámetro $\mu \in \mathbb{R}$ y cierta variable aleatoria X .

- ① Probar que si $g : \mathbb{R} \rightarrow \mathbb{R}$ es una función derivable en μ entonces $a_n(g(X_n) - g(\mu)) \xrightarrow{\mathcal{D}} g'(\mu)X$.

Sugerencia: Observar que para todo $x \in \mathbb{R}$ tenemos la escritura

$$g(x) = g(\mu) + (g'(\mu) + T(x))(x - \mu)$$

donde $T : \mathbb{R} \rightarrow \mathbb{R}$ es una función continua con $T(\mu) = 0$.

- ② Probar que si $g : \mathbb{R} \rightarrow \mathbb{R}$ es una función con derivada segunda en μ y tal que $g'(\mu) = 0$ entonces $a_n^2(g(X_n) - g(\mu)) \xrightarrow{\mathcal{D}} \frac{g''(\mu)}{2}X^2$.

Ejercicio 8.6

- a) Sean $(X_n)_{n \geq 1}$ variables aleatorias i.i.d. con distribución uniforme en el intervalo $[0, \theta]$: $X_i \sim \mathcal{U}[0, \theta]$. Demostrar que

$$\sqrt{n}(2\bar{X}_n - \theta) \xrightarrow{\mathcal{D}} W \quad \text{con } W \sim \mathcal{N}(0, \theta^2/3)$$

- b) Usando el ítem anterior y el método Delta, probar que

$$Y_n = \sqrt{n} [\ln(2\bar{X}_n) - \ln \theta] \xrightarrow{\mathcal{D}} V \quad \text{con } V \sim \mathcal{N}(0, 1/3).$$

Proposición 8.3

Sea $(X_n, Y_n)_{n \geq 1}$, una sucesión de vectores aleatorios tal que para todo n las variables aleatorias X_n e Y_n son independientes, y satisfacen $X_n \xrightarrow{\mathcal{D}} X$ e $Y_n \xrightarrow{\mathcal{D}} Y$. Entonces

$$X_n + Y_n \xrightarrow{\mathcal{D}} X_0 + Y_0,$$

donde X_0 e Y_0 son independientes y tales que $X_0 \sim X$ e $Y_0 \sim Y$.

dem: Por el Teo 8.8 i) tenemos

$$\begin{aligned} X_n &\xrightarrow{\mathcal{D}} X \rightarrow \varphi_{X_n}(t) \xrightarrow{t \rightarrow \infty} \varphi_X(t) \quad \forall t \in \mathbb{R} \\ Y_n &\xrightarrow{\mathcal{D}} Y \rightarrow \varphi_{Y_n}(t) \rightarrow \varphi_Y(t) \end{aligned}$$

Además, como x_n e y_n son independientes, x el
Teo 8.3 tenemos

$$\begin{aligned}\varphi_{x_n+y_n}(t) &= \varphi_{x_n}(t) \cdot \varphi_{y_n}(t) \xrightarrow{n \rightarrow \infty} \varphi_x(t) \varphi_y(t) = \\ &= \varphi_{x_0}(t) \varphi_{y_0}(t) = \varphi_{x_0+y_0}(t)\end{aligned}$$

\uparrow
pues $x_0 \sim x$

\uparrow
Teo 8.3

Luego

$$\varphi_{x_n+y_n}(t) \xrightarrow{n \rightarrow \infty} \varphi_{x_0+y_0}(t) \quad \forall t \in \mathbb{R}.$$

Para usar la parte ii) del Teo 8.8, necesitamos ver que $\varphi_{x_0+y_0}$ es continua en 0. Como $\varphi_{x_0+y_0}$ es la función característica de la var x_0+y_0 resulta uniformemente continua en \mathbb{R} (Lema 8.2), en particular cont en 0.

entonces $x_n+y_n \xrightarrow{\mathcal{D}} x_0+y_0$ ✓

9. Esperanza y Distribución condicional

Probabilidades y Estadística (M)

María Eugenia Szretter Noste

Departamento de Matemática e
Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Primer cuatrimestre 2020



Predictión

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad y $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ (que notaremos directamente \mathcal{L}^2) el espacio de variables aleatorias X definidas en Ω con $E[X^2] < +\infty$, que definimos cuando hablamos de esperanzas.

Sea Y una v.a. que no podemos observar, cuyo valor queremos predecir a partir de otras v.a.: X_1, \dots, X_n . Por ejemplo, Y puede ser el valor de una variable en el futuro, como una variable climática o bursátil, o una variable imposible de medir directamente como el estado del núcleo de un reactor nuclear, o la posición de un cuerpo celeste.

El problema de predecir una variable Y no observable a partir de observar otras X_1, \dots, X_n puede resumirse a buscar una función $h(X_1, \dots, X_n)$ que esté cerca de Y en algún sentido. Escribimos $\hat{Y} = h(X_1, \dots, X_n)$ y lo llamamos un predictor de Y .

Predictión

Primero necesitamos especificar el conjunto \mathcal{P} de predictores admisibles. Usualmente se toman todas las funciones tales que $h(X_1, \dots, X_n) \in \mathcal{L}^2$, pero a veces se restringe más la clase para facilitar el cálculo. Por ejemplo, permitiendo sólo funciones lineales de las v.a. observables.

En segundo lugar, tenemos que definir un criterio de optimalidad que nos ayude a identificar el “mejor” predictor.

Definición 9.1 (error cuadrático medio)

Dada una variable aleatoria Y y otra variable aleatoria \hat{Y} llamada predictor para Y , definimos el **error cuadrático medio (ECM)** que cometemos al predecir a Y con \hat{Y} mediante la fórmula

$$ECM(Y, \hat{Y}) = E[(Y - \hat{Y})^2].$$

Predictión

Definición 9.2

Dada una clase de predictores \mathcal{P} para la variable Y , diremos que \hat{Y}_* es el **predictor óptimo** (o de mínimo error cuadrático medio) en la clase si $\hat{Y}_* \in \mathcal{P}$ y además

$$ECM(Y, \hat{Y}) \geq ECM(Y, \hat{Y}_*) , \forall \hat{Y} \in \mathcal{P} .$$

Es decir, \hat{Y}_* es el elemento de la clase \mathcal{P} con el que menos error cometemos para predecir a Y . Lo llamaremos “mejor predictor de Y en la clase \mathcal{P} ”.

Vimos que $\langle X, Y \rangle = E[XY]$ define un producto interno en $\mathcal{L}^2(\Omega, \mathcal{F}, P)$. El producto interno define una norma en $\mathcal{L}^2(\Omega, \mathcal{F}, P)$, $\|X\| = \sqrt{E(X^2)}$. Queda entonces definida la distancia

$$d(X, Y) = \|X - Y\| = \sqrt{E[(X - Y)^2]}.$$

Luego, el $ECM(Y, \hat{Y}) = \|Y - \hat{Y}\|^2$. Identificamos dos variables en \mathcal{L}^2 , decimos $X = Y$ siempre que $P(X = Y) = 1$. También se escribe $X = Y$ (c.s.).

Definición 9.3

Decimos que las variables aleatorias X e Y en \mathcal{L}^2 son *ortogonales* si $\langle X, Y \rangle = E(XY) = 0$.

$\mathcal{P} \subset \mathcal{L}^2$ es un subespacio vectorial si $a\hat{Y}_1 + b\hat{Y}_2 \in \mathcal{P}$ para cualquier $a, b \in \mathbb{R}$ y $\hat{Y}_1, \hat{Y}_2 \in \mathcal{P}$.

Teorema 9.1

Sea \mathcal{P} un subespacio vectorial. Si $E[(Y - \hat{Y}_*)\hat{Y}] = 0 \forall \hat{Y} \in \mathcal{P}$ entonces

$$E[(Y - \hat{Y}_*)^2] \leq E[(Y - \hat{Y})^2], \forall \hat{Y} \in \mathcal{P}.$$

Luego, el error de predicción $Y - \hat{Y}_*$ es ortogonal a todo elemento de \mathcal{P} .

demos: Sea $\hat{Y} \in \mathcal{P}$. Entonces:

$$\begin{aligned} \text{ECM}(\hat{Y}, y) &= E((y - \hat{Y})^2) = E((y - \hat{Y}_*) + (\hat{Y}_* - \hat{Y}))^2 \\ &= E((y - \hat{Y}_*)^2) + E((\hat{Y}_* - \hat{Y})^2) + \\ &\quad + 2E((y - \hat{Y}_*)(\hat{Y}_* - \hat{Y})). \end{aligned}$$

Como $\hat{Y}_* - \hat{Y} \in \mathcal{P}$, x la cond de ortogonalidad tenemos $E((Y - \hat{Y}_*)(\hat{Y}_* - \hat{Y})) = 0$.

$$\begin{aligned}\therefore ECM(\hat{Y}, Y) &= E((Y - \hat{Y}_*)^2) + E((\hat{Y}_* - \hat{Y})^2) \\ &\geq E((Y - \hat{Y}_*)^2) = ECM(Y, \hat{Y}_*).\end{aligned}$$

y por lo tanto \hat{Y}_* es óptimo.

Teorema 9.2

Sea \mathcal{P} un subespacio vectorial de dimensión finita generado por

$\hat{Y}_1, \dots, \hat{Y}_k$:

$\mathcal{P} = \{\sum_{i=1}^k a_i \hat{Y}_i : a_i \in \mathbb{R}\}$. Entonces

$$E[(Y - \hat{Y}_*) \hat{Y}_i] = 0, 1 \leq i \leq k \Rightarrow E[(Y - \hat{Y}_*)^2] \leq E[(Y - \hat{Y})^2], \forall \hat{Y} \in \mathcal{P}.$$

dem:

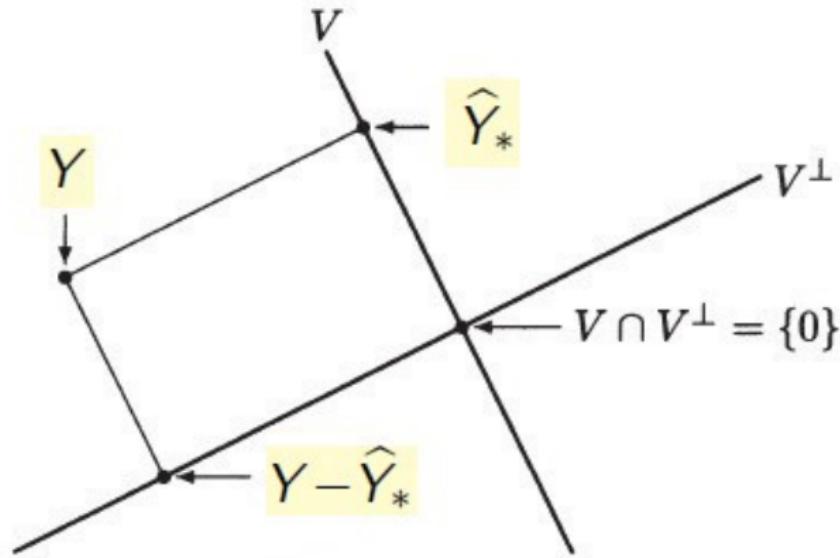
$\hat{Y} \in \mathcal{P}$, luego $\exists a_1, \dots, a_k \in \mathbb{R}$ tales que

$$\hat{Y} = \sum_{i=1}^k a_i \hat{Y}_i$$

$$\begin{aligned} E((Y - \hat{Y}_*) \hat{Y}) &= E\left((Y - \hat{Y}_*) \left[\sum_{i=1}^k a_i \hat{Y}_i\right]\right) \\ &= \sum_{i=1}^k a_i E((Y - \hat{Y}_*) \hat{Y}_i) = 0 \quad \text{por hipótesis} \end{aligned}$$

Luego \hat{Y}_* cumple las hipótesis del Teo anterior ✓

Figura 1: Geometría de la descomposición ortogonal y proyecciones



Predictores óptimos

Comenzaremos buscando predictores óptimos en diferentes clases. Para ello, pensemos en el vector aleatorio (X, Y) y buscaremos predictores óptimos para Y basados en X . Consideremos las siguientes clases:

$$\mathcal{P}_0 = \{a : a \in \mathbb{R}\}$$

$$\mathcal{P}_1 = \{bX + a : a, b \in \mathbb{R}\}$$

$$\mathcal{P}_2 = \{cX^2 + bX + a : a, b, c \in \mathbb{R}\}$$

¿Cómo encontrar el predictor óptimo en cada clase?

Predictores óptimos: mejor predictor constante

Opción 1: En cada caso tenemos generadores del espacio \mathcal{P} . Luego, basta garantizar ortogonalidad respecto de cada uno de los generadores.

Opción 2: Basta considerar una función que a cada predictor le asigna la pérdida correspondiente y luego minimizar dicha función.

Por ejemplo, para buscar el predictor óptimo en la clase \mathcal{P}_0 debemos encontrar a_0 que minimice

$$H(a) = E[(Y - a)^2].$$

Pero este problema ya lo resolvimos en la Proposición 6.3. Sabemos que $a_0 = \mu_Y = E[Y]$. ¿Cuál es el precio que pagamos al predecir a Y por su esperanza? En otras palabras, ¿cuánto vale $H(a_0)$? También lo vimos,

$$H(a_0) = E[(Y - a_0)^2] = E[(Y - \mu_Y)^2] = V(Y).$$

A μ_Y se lo denomina el mejor predictor constante de la variable Y .

Predictores óptimos: mejor predictor constante

En la práctica se usa un predictor constante solamente cuando no se cuenta con otras variables que puedan ayudar en la predicción.

Este resultado refuerza la idea con la que nos encontramos antes: las variables que más cuesta predecir con una constante (por las que pagamos un mayor precio al usar una constante como predicción) son aquellas que tienen mayor varianza, es decir, las más dispersas. La varianza es justamente el precio que pagamos al cambiar a una variable por su esperanza (que es el mejor resumen constante en el sentido del ECM).

Óptimo lineal

En cambio, para encontrar el mejor predictor en la clase \mathcal{P}_1 , buscamos a_0, b_0 que minimicen la función

$$H(a, b) = E[(Y - (bX + a))^2]$$

Una base de \mathcal{P}_1 está dada por $\{1, X\}$. ($X \neq c$
const.)

Luego \hat{Y}_* el predictor óptimo en la clase \mathcal{P}_1
debe satisfacer:

$$\hat{Y}_* = b_0 X + a_0.$$

$$E((Y - b_0 X - a_0)X) = 0 \quad (1)$$

$$E((Y - b_0 X - a_0)1) = 0 \quad (2)$$

$$E((Y - b_0x - a_0)x) = 0 \quad (1)$$

(Copiamos)

$$E((Y - b_0x - a_0)1) = 0 \quad (2)$$

De (2): $E(Y) - b_0E(x) - a_0 = 0 \Leftrightarrow a_0 = E(Y) - b_0E(x)$

De (1): $E((Y - b_0x - E(Y) + b_0E(x))x) = 0$

$\Leftrightarrow E\left(\left([Y - E(Y)] - b_0[x - E(x)]\right)x\right) = 0$

$$E([Y - E(Y)]x) = b_0E([x - E(x)]x) \quad (3)$$

Observemos que:

$$\begin{aligned} E([x - E(x)]x) &= E(x^2 - E(x)x) \\ &= E(x^2) - E(x)E(x) = V(x) > 0 \end{aligned}$$

$$\begin{aligned} E([y - E(y)]x) &= E(xy - E(y)x) \\ &= E(xy) - E(y)E(x) = \text{cov}(xy) \end{aligned}$$

Luego, de (3)

$$b_0 = \frac{\text{cov}(x, y)}{V(x)}, \quad a_0 = E(y) - \frac{\text{cov}(x, y)E(x)}{V(y)}$$

Obtenemos que H se minimiza en el par (a_0, b_0) siendo

$$b_0 = \frac{\text{Cov}(X, Y)}{V(X)}, \quad a_0 = E[Y] - b_0 E[X].$$

A $b_0 X + a_0$ se lo denomina el

mejor predictor lineal de la variable Y basado en X . Si recordamos que el coeficiente de correlación entre las variables aleatorias X e Y está dado por

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}},$$

tenemos que el mejor predictor lineal para Y basado en X está dado por

$$\rho \frac{\sigma_X}{\sigma_Y} X + (\mu_Y - \rho \frac{\sigma_X}{\sigma_Y} \mu_X) = \rho \frac{\sigma_Y}{\sigma_X} (X - \mu_X) + \mu_Y$$

donde $\sigma_X^2 = V(X)$, $\mu_X = E(X)$, $\sigma_Y^2 = V(Y)$, $\mu_Y = E(Y)$ y $\rho = \rho(X, Y)$.

¿Cuál es el precio que pagamos al predecir a Y con el mejor predictor lineal basado en X ?

$$\begin{aligned}
 H(a_0, b_0) &= E((Y - b_0 X - a_0)^2) = E((Y - E(Y)) - b_0(X - E(X))^2) \\
 &= E([Y - E(Y)]^2 + b_0^2 [X - E(X)]^2 - 2b_0 [X - E(X)][Y - E(Y)]) \\
 &= V(Y) + b_0^2 V(X) - 2b_0 \text{cov}(X, Y) \\
 &\stackrel{\uparrow}{=} V(Y) + \frac{\text{cov}^2(X, Y)}{V(X)} - 2 \frac{\text{cov}^2(X, Y)}{V(X)} = V(Y) - \frac{\text{cov}^2(X, Y)}{V(X)V(Y)}V(Y) \\
 b_0 &= \frac{\text{cov}(X, Y)}{V(X)} \quad \rho^2 = \frac{\text{cov}^2(X, Y)}{V(X)V(Y)} \\
 &= V(Y) [1 - \rho^2]
 \end{aligned}$$

$$H(a_0, b_0) = E[(Y - b_0 X - a_0)^2] = \sigma_Y^2(1 - \rho^2).$$

$$H(a_0, b_0) = E[(Y - b_0X - a_0)^2] = \sigma_Y^2(1 - \rho^2).$$

Siendo que $H(a_0, b_0) \geq 0$ y que $\sigma_Y^2 \geq 0$, obtenemos que Recordemos que

$$-1 \leq \rho(X, Y) \leq 1$$

Por lo que, si $\rho^2 = 1$, tenemos que $H(a_0, b_0) = 0$, en cuyo caso

$$Y = b_0X + a_0$$

Además,

$$\rho = 1 \rightarrow b_0 > 0$$

$$\rho = -1 \rightarrow b_0 < 0$$

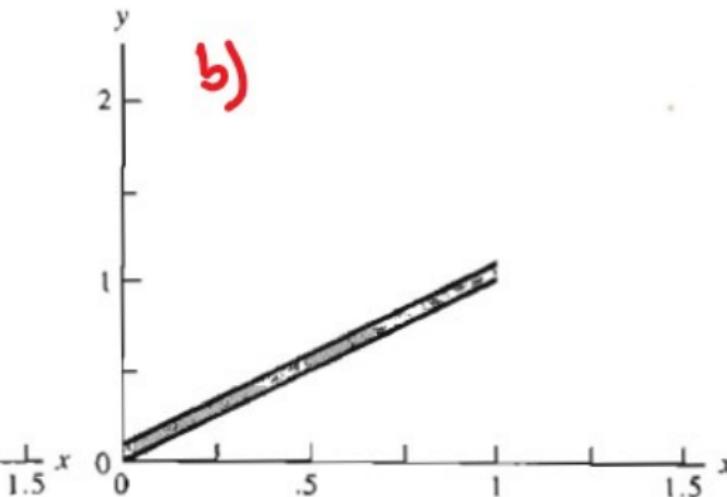
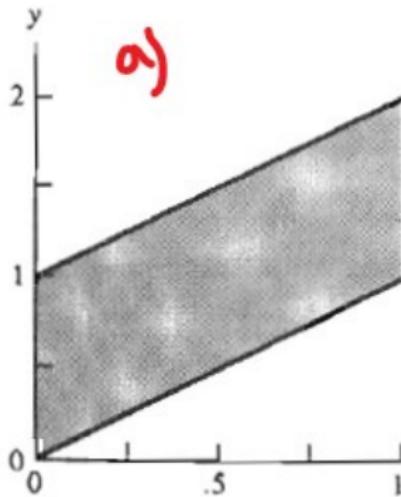
En general, cuanto más cerca esta $|\rho|$ de uno, menor es el precio que pagamos al predecir linealmente.

Ejemplo 9.1 (Dos vectores uniformes)

Veamos dos ejemplos de un vector $(X, Y) \sim \mathcal{U}(R)$, cuyos soportes están en la Figura 2

Figura 2: Soporte de dos vectores uniformes, $(X, Y) \sim \mathcal{U}(R)$ con a)
 $R = \{(x, y) : 0 \leq x \leq 1, x \leq y \leq x + 1\}$ con $\rho = 1/\sqrt{2} = 0,707$. Como
 $V(Y) = \frac{1}{6}$ resulta $H(a_0, b_0) = \frac{1}{12} = 0,08333$ y en b)

$R = \{(x, y) : 0 \leq x \leq 1, x \leq y \leq x + \frac{1}{10}\}$ con $\rho = \frac{\sqrt{100}}{\sqrt{101}} = 0,995$. Acá
 $V(Y) = \frac{1}{12} + \frac{1}{120}$ entonces $H(a_0, b_0) = \frac{1}{1200} = 0,0008333$. (Fuente: Casella, G,
Berger R. *Statistical Inference*, (2002), ejemplos 4.5.4 y 4.5.8)



$E(X)$	$V(X)$	$E(Y)$	$V(Y)$	$E(XY)$	Cov(X, Y)
--------	--------	--------	--------	---------	---------------

a) $\frac{1}{2}$ $\frac{1}{12}$ 1 $\frac{1}{6}$ $\frac{7}{12}$ $\frac{1}{12}$

b) $\frac{1}{2}$ $\frac{1}{12}$ $\frac{11}{20}$ $\frac{1}{12} + \frac{1}{120}$ $\frac{43}{120}$ $\frac{1}{12}$

a) $\widehat{Y}_{*,\text{opt lin}} = X + \frac{1}{2}$, $\rho = \frac{1}{\sqrt{2}} = 0,707$ y precio $H(a_0, b_0) = \frac{1}{12} = 0,08333$

b) $\widehat{Y}_{*,\text{opt lin}} = X + \frac{1}{20}$, $\rho = \frac{\sqrt{100}}{\sqrt{101}} = 0,995$ y precio
 $H(a_0, b_0) = \frac{1}{1200} = 0,0008333$

En ambos casos hay una relación linealmente creciente entre X e Y . La covarianza entre ambos es la misma, $\frac{1}{12}$. En el caso (b) sin embargo, el óptimo lineal representa un mejor resumen del vínculo entre X e Y que en (a). Eso se ve reflejado en el menor precio $H(a_0, b_0)$.

Predictión, caso X variable discreta

Podemos seguir buscando predictores óptimos en diferentes clases. Pero surge la siguiente pregunta: Si soy libre de hacer con X cualquier tipo de cuenta, ¿cuál es la mejor que puedo hacer para predecir a Y ? Es decir, buscamos predictor óptimo para Y en la clase

$$\mathcal{P} = \{g(X), g : \mathbb{R} \rightarrow \mathbb{R} \text{ boreiana}\}.$$

Resolvamos este problema asumiendo que X es un vector discreto, de rango finito. Supongamos que $R_X = \{x_1, \dots, x_k\}$. En tal caso,

$$\mathcal{P} = \left\{ \sum_{i=1}^k a_i I_{\{X=x_i\}} : a_i \in \mathbb{R} \right\} \quad (a_i = g(x_i))$$

En otras palabras, cuando la variable X es discreta, la clase \mathcal{P} es un espacio vectorial finito dimensional, generado por las funciones indicadoras de cada uno de los valores que la variable X toma. Es decir, \mathcal{P} está generado por $I_{\{X=x_i\}}$ para $1 \leq i \leq k$.

Predictión, caso X variable discreta

Luego, \hat{Y}_* debe ser de la forma $\hat{Y}_* = \sum_{i=1}^k a_i^* I_{\{X=x_i\}}$ para valores a_i^* de forma tal de verificar

$$E[(Y - \hat{Y}_*) I_{\{X=x_i\}}] = 0, \quad 1 \leq i \leq k.$$

$$\begin{aligned} E[(Y - \hat{Y}_*) I_{\{X=x_i\}}] &= 0 \Leftrightarrow \\ E(Y I_{\{X=x_i\}}) &= E(\hat{Y}_* I_{\{X=x_i\}}) \quad \text{son todos cero excepto } j=i \\ &= E\left(\sum_{j=1}^k a_j I_{\{X=x_j\}} I_{\{X=x_i\}}\right) \\ &= \sum_{j=1}^k a_j E(I_{\{X=x_j\}} I_{\{X=x_i\}}) \\ &= a_i P(X=x_i) = a_i p_X(x_i) \end{aligned}$$

$\Rightarrow a_i = \frac{E(Y I_{\{X=x_i\}})}{p_X(x_i)}$

Tenemos así que la solución del sistema está dada por

$$a_i^* = \frac{E[YI_{\{X=x_i\}}]}{P(X=x_i)} = \frac{E[YI_{\{x_i\}}(X)]}{P(X=x_i)}$$

El numerador es la esperanza de una función $g(X, Y) = YI_{\{x_i\}}(X)$.

Cuando el vector (X, Y) es discreto, tenemos que

$$\begin{aligned} a_i^* &= \frac{E[YI_{\{x_i\}}(X)]}{P(X=x_i)} = \frac{E[g(X, Y)]}{P(X=x_i)} = \frac{\sum_{x,y} g(x, y)p_{XY}(x, y)}{p_X(x_i)} \\ &= \frac{\sum_{x,y} yI_{\{x_i\}}(x)p_{XY}(x, y)}{p_X(x_i)} = \frac{\sum_y y p_{XY}(x_i, y)}{p_X(x_i)} = \sum_y y \frac{p_{XY}(x_i, y)}{p_X(x_i)}. \end{aligned}$$

Notemos que a_i^* resulta ser una esperanza respecto de una nueva distribución. Más precisamente es la esperanza de Y en un nuevo mundo donde asumimos que $X = x_i$.

Cuando el vector (X, Y) es discreto,

$$a_i^* = \sum_y y \frac{p_{XY}(x_i, y)}{p_X(x_i)}$$

En tal caso, la función de probabilidad puntual de Y en este **nuevo mundo** está dada por $\frac{p_{XY}(x_i, y)}{p_X(x_i)}$ y entonces a_i^* resulta ser la esperanza de Y con esta función de probabilidad puntual.

Más generalmente, si conocemos la distribución condicional de $Y|X = x_i$, tenemos entonces que

$$a_i^* = \frac{E[YI_{\{X=x_i\}}]}{P(X = x_i)}$$

se calcula utilizando la distribución condicional.

Entonces, vamos a definir la distribución condicional de $Y|X$ y a partir de ella, calcularemos los predictores óptimos en los distintos casos.

Predictión, caso general

Proposición 9.1

Dado el vector aleatorio (X, Y) con $X, Y \in \mathcal{L}^2$ existe a lo sumo una única función $g^*(X) \in \mathcal{L}^2$ verificando

$$E[(Y - g^*(X))h(X)] = 0 \quad \text{para toda } h(X) \in \mathcal{L}^2 \quad (1)$$

Es decir, si g_1^* y g_2^* satisfacen la condición (1), entonces $P(g_1^*(X) = g_2^*(X)) = 1$

Demostración.

Sean g_1^* y g_2^* satisfaciendo la condición (1). Tenemos entonces que

$E[g_i^*(X)h(X)] = E[Yh(X)]$ para $i = 1, 2$. Restando estas dos identidades, obtenemos que

$$E[(g_1^*(X) - g_2^*(X))h(X)] = 0 \quad \forall h(X) \in \mathcal{L}^2$$

En particular, eligiendo $h(X) = \text{signo}(g_1^*(X) - g_2^*(X))$ concluimos que

$E[|g_1^*(X) - g_2^*(X)|] = 0$ de donde deducimos que $P(g_1^*(X) - g_2^*(X) = 0) = 1$ y por consiguiente $P(g_1^*(X) = g_2^*(X)) = 1$.

La Proposición 9.1 también vale si en (1) se toman todas las funciones h borelianasy acotadas.

Esperanza Condicional

Teorema 9.3

Dado el vector aleatorio (X, Y) con $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P) = \mathcal{L}^2$ existe una única $g^*(X) \in \mathcal{L}^2$ satisfaciendo

$$E[(Y - g^*(X))h(X)] = 0, \quad \text{para toda } h(X) \in \mathcal{L}^2 \quad (2)$$

Demostración.

La **existencia** surge de resultados de teoría de la medida. No la probamos en general en esta materia. La probaremos en casos particulares. De hecho, en esos casos daremos una construcción explícita.

La **unicidad** es consecuencia de la Proposición 9.1. □

Definición 9.4 (esperanza condicional)

Dado el vector aleatorio (X, Y) con $X, Y \in \mathcal{L}^2$ definimos **la esperanza condicional de Y dado X** como la única función de X satisfaciendo (2).

Esperanza condicional

Notación: Comúnmente, se utiliza $E[Y | X]$ en lugar de $g^*(X)$.

Denotaremos por $E[Y | X = x]$ al valor que toma la variable aleatoria $E[Y | X]$ cuando $X = x$.

Concretamente, para determinar que una variable aleatoria candidata \mathbf{W} es la esperanza condicional de Y dado X debemos verificar que:

- dicha candidata \mathbf{W} sea función de X , es decir, $\mathbf{W} = g^*(X)$ (con g^* boreiana),
- cumpla $E[(Y - g^*(X))h(X)] = 0$, para toda $h(X) \in \mathcal{L}^2$

Esta definición es muy general pero no nos dice nada de cómo calcularla. Veamos algunas propiedades de la esperanza condicional, y luego nos concentraremos en obtenerla en los distintos casos.

Teorema 9.4

Sean $X, Y \in \mathcal{L}^2$. Entonces podemos considerar el subespacio $\mathcal{S} \subset \mathcal{L}^2$,

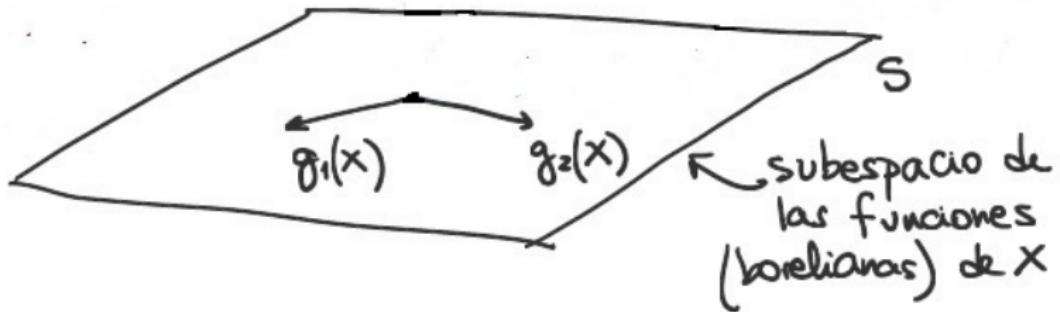
$$\mathcal{S} = \left\{ g(X), \text{ con } g : \mathbb{R} \rightarrow \mathbb{R} \text{ boreiana con } E([g(X)]^2) < \infty \right\}$$

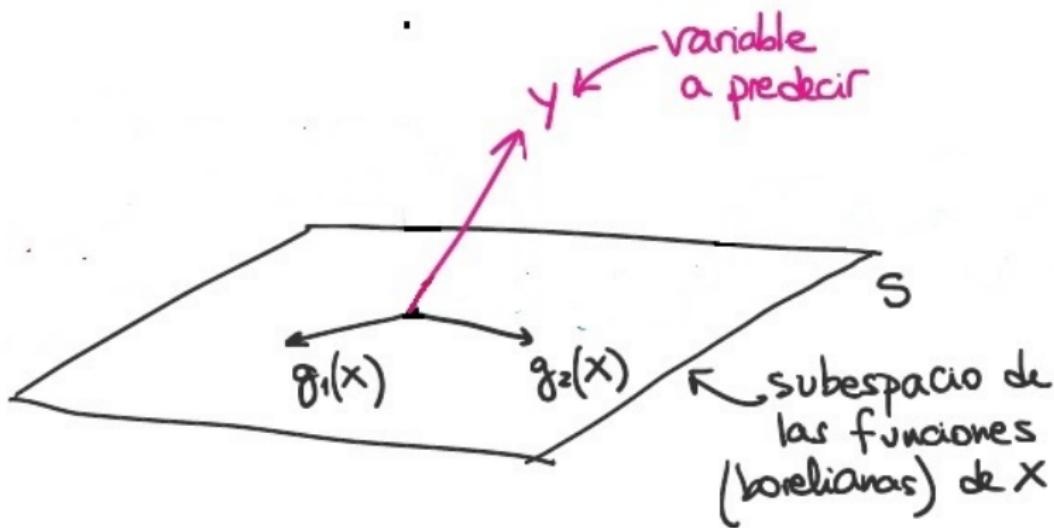
Entonces la variable $g^*(X) = E[Y | X]$ es la proyección de Y sobre \mathcal{S} . Es decir, es el elemento de \mathcal{S} que cumple

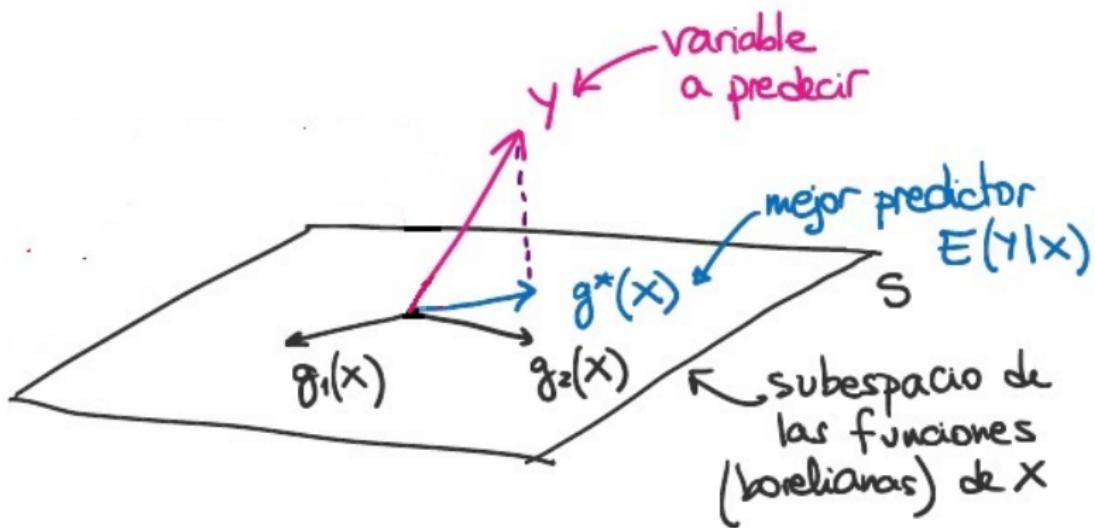
$$E[(Y - g^*(X))^2] \leq E[(Y - g(X))^2]$$

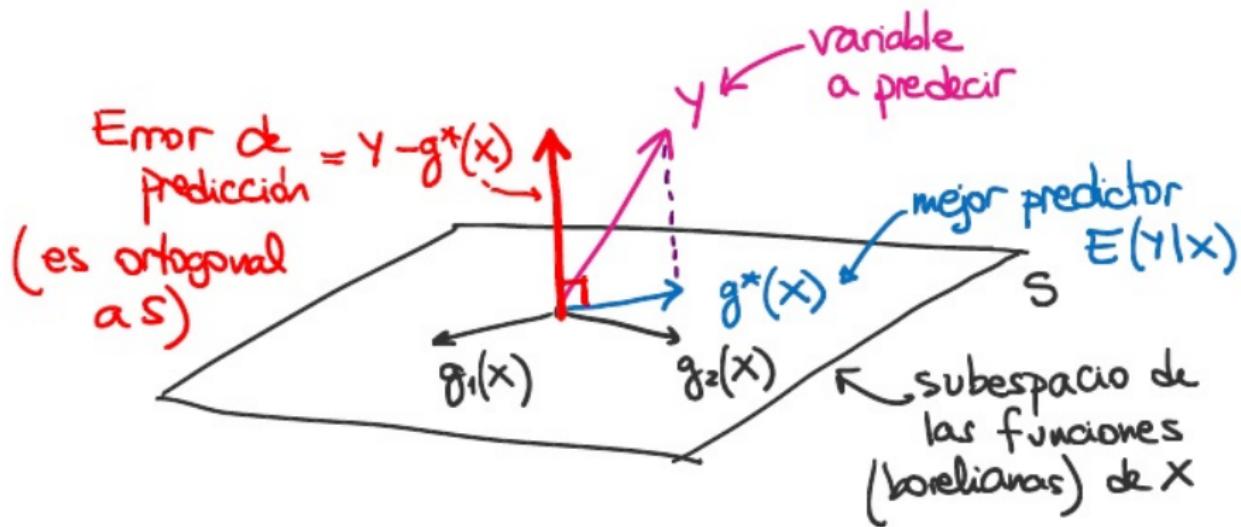
para toda $g(X) \in \mathcal{S}$.

Moraleja: La esperanza condicional de Y dado X resulta ser la mejor función de X para predecir a Y .









Demostración.

Por definición de $E[Y | X]$ sabemos que $Y - g^*(X)$ es ortogonal a toda $g(X) \in \mathcal{S}$. Luego, por Teorema 9.1 resulta que

$$E[(Y - g^*(X))^2] \leq E[(Y - g(X))^2], \quad \forall g(X) \in \mathcal{S}. \quad (3)$$

□

Teorema 9.5 (propiedades de la esperanza condicional)

Dado el vector aleatorio (X, Y) , tenemos que

1) $E[r(X)s(Y) | X] = r(X) E[s(Y) | X]$

1) Basta ver que $r(x)E(s(y)|x)$ cumple la definición de $E(r(x)s(y)|x)$, es decir, que para toda $h(x) \in \mathcal{L}^2$ se tiene:

$$E([r(x)s(y) - r(x)E(s(y)|x)]h(x)) = 0$$

Pero $E([r(x)s(y) - r(x)E(s(y)|x)]h(x)) =$

$$= E\left(\underbrace{r(x)h(x)}_{\text{función de } X} [s(y) - E(s(y)|x)]\right) = 0$$

por definición de $E(s(y)|x)$.

Además, claramente $r(x)E(s(y)|x)$ es función de x .

Toerema 9.5 (cont.)

Dado el vector aleatorio (X, Y) , tenemos que

2) $E[Y | X] = E[Y]$ si (X, Y) son independientes.

2) Basta ver que $E(Y)$ satisface las propiedades requeridas. •) $E(Y)$ es una función (constante) de X .

•) $E((Y - E(Y)) \cdot h(X)) = E(Yh(X) - E(Y)h(X))$

Candidato
a esperanza condicional

$$\stackrel{\uparrow}{=} E(Y) \cdot E(h(X)) - E(Y) E(h(X)) = 0 \quad \checkmark$$

por indep.

entre x e y en el 1^{er} término, y \times linealidad en el 2^o

Teorema 9.5 (cont.)

Dado el vector aleatorio (X, Y) , tenemos que

3) $E[E[Y | X]] = E[Y]$

3) Sabemos que $E([Y - E(Y|X)] h(x)) = 0 \quad \forall h(x) \in \mathcal{L}^2$

En particular, si tomamos $h(x) = 1$ tenemos:

$$E(Y) = E(E(Y|X)) \quad \checkmark$$

4) $E[r(X) | X] = r(X)$

4) Queremos
ver que

•) $r(x)$ es función de x .

•) Claramente $E[(r(x) - r(x)) h(x)] = 0 \quad \forall h(x) \in \mathcal{L}^2 \quad \checkmark$

Teorema 9.5 (cont.)

Dado el vector aleatorio (X, Y) , tenemos que

5) Linealidad: $E[aY_1 + bY_2 | X] = aE[Y_1 | X] + bE[Y_2 | X]$.

$$\begin{aligned} & 5) E[(aY_1 + bY_2 - \underbrace{(aE(Y_1|X) + bE(Y_2|X))}_{\substack{\text{candidata a esperanza} \\ \text{condicional}}}) h(x)] \\ & \stackrel{\substack{\uparrow \\ \text{linealidad}}}{=} a E[(Y_1 - E(Y_1|X)) h(x)] + b E[(Y_2 - E(Y_2|X)) h(x)] = 0 \quad \checkmark \end{aligned}$$

A continuación presentamos las distribuciones condicionales.

Distribución Condicional

Caso (X, Y) Discreto

Definición 9.5 (f.p.p. de Y condicional a $X = x$)

Dado el vector discreto (X, Y) para cada x con $p_X(x) > 0$ definimos una nueva función de probabilidad dada por

$$p_{Y|X=x}(y) = P(Y = y | X = x) = \frac{p_{XY}(x, y)}{p_X(x)}.$$

Se denomina la función de probabilidad puntual de Y condicional a $X = x$

También se la suele notar:

$$p_{Y|X}(y | x) = p_{Y|X=x}(y)$$

Distribucion Condicional (caso discreto)

Lema 9.6

Para cada x con $p_X(x) > 0$ tenemos que :

- ① $p_{Y|X=x}(\cdot)$ es una función de probabilidad puntual.
- ②

$$P(Y \in A | X = x) = \sum_{y \in A} p_{Y|X=x}(y)$$

- ③ Si X e Y son independientes, entonces $p_{Y|X=x}(\cdot) = p_Y(\cdot)$

1) Queremos ver que $\rightarrow p_{Y|X=x}(y) \geq 0 \quad \forall y \quad \checkmark$

$$\therefore \sum_{y \in R_Y} p_{Y|X=x}(y) = 1.$$

2) \checkmark (es la definición)

$$\sum_{y \in R_Y} p_{Y|X=x}(y) = \sum_{y \in R_Y} \frac{p_{XY}(y, x)}{p_X(x)} = \frac{p_X(x)}{p_X(x)} = 1 \quad \checkmark$$

$$3) p_{Y|X=x}(y) = \frac{p_{XY}(x, y)}{p_X(x)} = \frac{p_X(x) p_Y(y)}{p_X(x)} = p_Y(y) \quad \forall y \in R_Y, \forall x \in R_X$$

Esperanza Condicional (caso discreto)

Proposición 9.2

Dado el vector discreto (X, Y) , tenemos una fórmula explícita para calcular la esperanza condicional, dada por

$$E[Y | X = x_i] = \sum_{y \in R_Y} y p_{Y|X=x_i}(y).$$

Llamemos $g^*(x) = \sum_{y \in R_Y} y p_{Y|X=x}(y)$. Para probar este resultado tenemos que probar que la variable aleatoria definida por $g^*(X)$ cumple:

- es función de X ✓
- el error de predicción $Y - g^*(X)$ es ortogonal a toda variable $h(X)$.

Como la proposición que sigue tiene a este resultado como caso particular, probaremos directamente la Proposición 9.3.

Esperanza Condicional (caso discreto)

Proposición 9.3

Dado el vector discreto (X, Y) , vale que

$$E[r(Y) | X = x_i] = \sum_y r(y)p_{Y|X=x_i}(y)$$

mientras que

$$E[\varphi(X, Y) | X = x_i] = E[\varphi(x_i, Y) | X = x_i] = \sum_y \varphi(x_i, y)p_{Y|X=x_i}(y)$$

Demostración.

Claramente, $g^*(x) = \sum_y \varphi(x, y)p_{Y|X=x}(y)$ es función de x . ✓



Queremos probar que

$$E[(\varphi(x, y) - g^*(x)) h(x)] = 0.$$

$$\Leftrightarrow E[\varphi(x, y) h(x)] = E[g^*(x) h(x)]$$

Calculemos $E[g^*(x) h(x)] = \sum_{x_i \in R_X} g^*(x_i) h(x_i) p_X(x_i)$

$$= \sum_{x_i \in R_X} \left[\sum_{y \in R_Y} \varphi(x_i, y) p_{Y|X=x_i}(y) \right] h(x_i) p_X(x_i)$$

$$= \sum_{x_i \in R_X} \sum_{y \in R_Y} \varphi(x_i, y) \frac{p_{XY}(x_i, y)}{p_X(x_i)} h(x_i) p_X(x_i)$$

$$= \sum_{x_i \in R_X} \sum_{y \in R_Y} \varphi(x_i, y) h(x_i) p_{XY}(x_i, y) = E[h(x)\varphi(x, y)]$$

Distribucion Condicional

Caso (X, Y) continuo

Definición 9.6

Dado el vector continuo (X, Y) para cada x con $f_X(x) > 0$ definimos una nueva función de densidad, dada por

$$f_{Y|X=x}(y) = \frac{f_{XY}(x, y)}{f_X(x)}.$$

Se denomina la función de densidad de Y condicional a (o dada) $X = x$.

Lema 9.7

Para cada x con $f_X(x) > 0$ tenemos que :

1) $f_{Y|X=x}(\cdot)$ es una función de densidad.

2) Si X e Y son independientes, entonces $f_{Y|X=x}(\cdot) = f_Y(\cdot)$

1) •) $f_{Y|X=x}(y) \geq 0 \quad \forall y \in \mathbb{R}$

•) $\int_{-\infty}^{+\infty} f_{Y|X=x}(y) dy = \left[\int_{-\infty}^{+\infty} f_{XY}(x,y) dy \right] \frac{1}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1 \quad \checkmark$

2) X e Y indep,

$$f_{Y|X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y) \quad \forall y \in \mathbb{R} \quad \checkmark$$

Proposición 9.4

Sea

$$g^*(x) = \int_{-\infty}^{+\infty} y f_{Y|X=x}(y) dy .$$

Tenemos entonces que $E[Y | X] = g^*(X)$.

Demostración.

Siendo $g^*(x)$ función de x , para garantizar que $g^*(X)$ es la esperanza condicional de Y dado X resta verifica la condición (2), es decir, debemos demostrar que para toda $h(X)$, $E[(Y - g^*(X))h(X)] = 0$. Ahora bien,

$$\begin{aligned} E[g^*(X)h(X)] &= \int \int g^*(x)h(x) f_{XY}(x,y) dx dy \\ &= \int g^*(x)h(x) f_X(x) dx \\ &= \int \left(\int y f_{Y|X=x}(y) dy \right) h(x) f_X(x) dx \\ &= \int \int yh(x) f_{XY}(x,y) dx dy = E[Yh(X)] \end{aligned}$$

Proposición 9.5

Sea (X, Y) un vector continuo, tenemos entonces que
 $E[\varphi(X, Y) | X] = g^*(X)$, siendo $g^*(x)$ la función dada por

$$g^*(x) = \int \varphi(x, y) f_{Y|X=x}(y) dy .$$

La demostración queda como ejercicio.

Resumiendo, tenemos que

$$E[\varphi(X, Y) | X = x] = \begin{cases} \sum_y \varphi(x, y) p_{Y|X=x}(y) & \text{si el vector } (X, Y) \\ & \text{es discreto} \\ \int_{-\infty}^{\infty} \varphi(x, y) f_{Y|X=x}(y) dy & \text{si el vector } (X, Y) \\ & \text{es continuo.} \end{cases}$$

Más generalmente, si conocemos la distribución de $Y | X = x$, la fórmula explícita para hallar la $E[\varphi(X, Y) | X = x]$ es la esperanza bajo esa distribución de $\varphi(x, Y)$.

Observemos que entonces podemos calcular la esperanza condicional en casos más generales que cuando el vector (X, Y) es discreto o continuo.

Dos casos relevantes aparecen como aplicación del cálculo de $E[\varphi(X, Y) | X = x]$:

- a) La probabilidad condicional de un evento.
- b) La varianza condicional

Del mismo modo que la probabilidad de un evento que involucra a una variable aleatoria se puede escribir como la esperanza de una función indicadora, definimos

Definición 9.7 (probabilidad condicional)

Dado un vector aleatorio (X, Y) , notamos

$$P(Y \in A | X = x) = E[I_A(Y) | X = x].$$

*y la llamamos **probabilidad condicional de que $\{Y \in A\}$ cuando $X = x$.***

Varianza condicional

Definición 9.8 (varianza condicional)

Dado un vector aleatorio (X, Y) , definimos la varianza condicional de Y dado X como la variable aleatoria que cuando $X = x$ toma el valor

$$V(Y | X = x) = E\left[\left(Y - E[Y | X = x]\right)^2 | X = x\right] \quad (4)$$

Es decir, $V(Y | X = x)$ es la varianza correspondiente a la distribución condicional de $Y | X = x$. Por lo general, denotaremos por $V(Y | X)$ a la variable aleatoria que cuando $X = x$ toma el valor dado por la fórmula (4).

Varianza condicional

La varianza condicional admite las siguientes representaciones, según el vector sea discreto o continuo:

$$V(Y | X = x) = \begin{cases} \sum_y \left(y - E[Y | X = x] \right)^2 p_{Y|X=x}(y) & \text{(discreto)} \\ \int_{-\infty}^{\infty} \left(y - E[Y | X = x] \right)^2 f_{Y|X=x}(y) dy & \text{(continuo).} \end{cases}$$

Varianza condicional

Tal como acontece con la propia varianza, contamos con una fórmula reducida también para la varianza condicional, dada en el siguiente lema:

Lema 9.8 (Fórmula reducida para la varianza condicional)

$$V(Y | X = x) = E[Y^2 | X = x] - \left(E[Y | X = x] \right)^2.$$

Dicho de otra forma,

$$V(Y | X) = E[Y^2 | X] - \left(E[Y | X] \right)^2.$$

Además, varianzas y esperanzas condicionales se relacionan:

Lema 9.9

$$V(Y) = E[V(Y | X)] + V(E[Y | X]).$$

Ejemplo 9.2: Salarios EE.UU.

La encuesta *Current Population Survey* llevada a cabo por la oficina censal estadounidense (U.S. Census Bureau) y la oficina de censo laboral estadounidense (U.S. Bureau of Labor Statistics (BLS)) es la fuente principal de estadísticas laborales para la población de Estados Unidos. El libro Hansen, B. (2020) *Econometrics*

<https://www.ssc.wisc.edu/~bhansen/econometrics/> publica la muestra de 50.742 trabajadores no militares a tiempo completo de la edición 2009 de la encuesta. El dataset puede encontrarse en la misma url. Podemos considerar dos variables aleatorias: $X = \text{años de estudio}$ e $Y = \text{salario por hora (en dólares)}$, construida dividiendo el ingreso anual dividido por las horas trabajadas en el año. Una distribución conjunta de las variables discretizadas, tomando a los más de 50 mil trabajadores como si fueran la población, nos da la tabla que sigue.

Ejemplo 9.2

Cuadro 1: Las variables resultantes de la encuesta *Current Population Survey*, basada en 50742 trabajadores a tiempo completo en EE.UU. en marzo 2009.

		$X =$	años	de	estudio	
		5	12	15	18.5	$p_Y(k)$
$Y =$ salario por hora	7	0.0094	0.0766	0.0516	0.0038	0.1415
	15	0.0077	0.1680	0.2028	0.0211	0.3996
	24	0.0010	0.0579	0.1357	0.0341	0.2287
	32	0.0001	0.0123	0.0454	0.0143	0.0721
	40	0.0002	0.0106	0.0507	0.0244	0.0859
	77	0.0000	0.0049	0.0361	0.0313	0.0722
	$p_X(j)$	0.0183	0.3304	0.5223	0.1290	1

El salario medio, $E(Y) = 23.8$ (redondeando decimales), lo cual sale de esta tabla. Pero si uno se pregunta cual es el salario medio o esperado para un trabajador que tiene 5 años de estudio, tiene que calcular

$$p_{Y|X=5}(k) = \frac{P(X=5, Y=k)}{p_X(5)} = \frac{p_{XY}(5,k)}{p_X(5)}$$

$$p_{Y|X=5}(k) = \frac{P(X=5, Y=k)}{p_X(5)} = \frac{p_{XY}(5,k)}{p_X(5)}$$

k	7	15	24	32	40	77
$p_{Y X=5}(k)$	0.5137	0.4232	0.0538	0.0032	0.0086	0.0000

$$p_{Y|X=15}(k) = \frac{P(X=15, Y=k)}{p_X(15)} = \frac{p_{XY}(15,k)}{p_X(15)}$$

	7	15	24	32	40	77
$p_{Y X=15}(k)$	0.0988	0.3882	0.2598	0.0869	0.0972	0.0691

¿Cómo lo visualizamos?

Figura 3: Función de p.p. condic. del salario, $p_{Y|X=j}(.)$, una curva para cada j nivel educativo

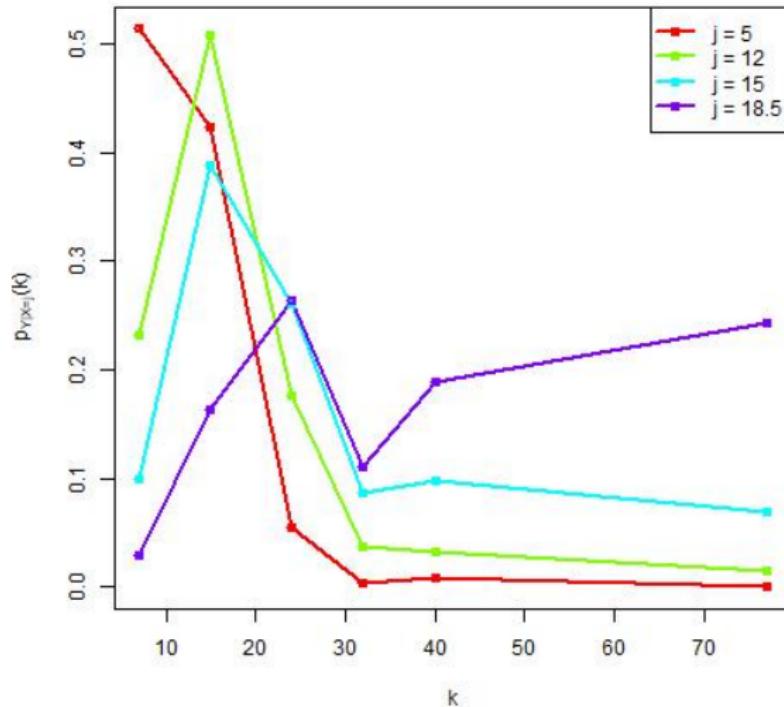
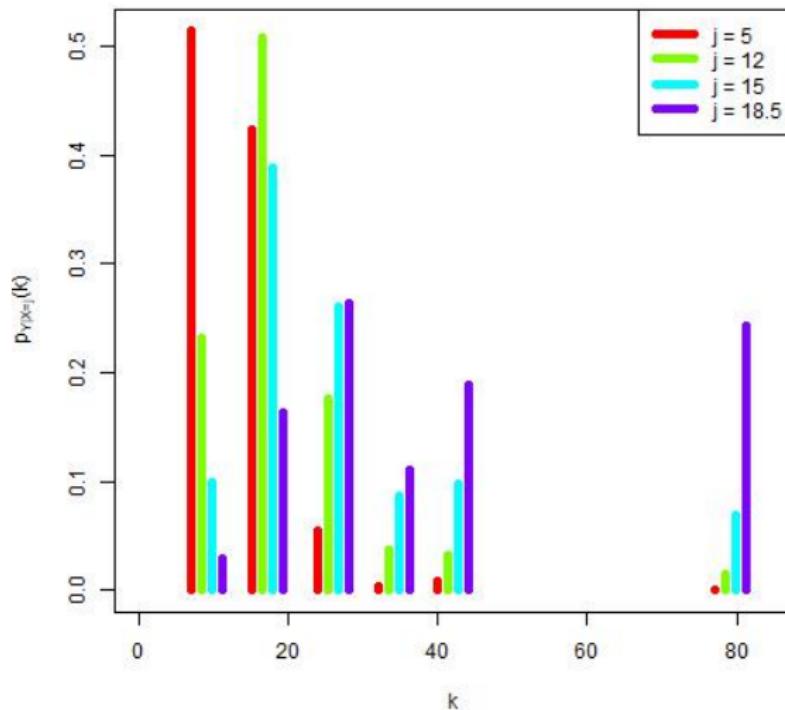


Figura 4: Función de probabilidad puntual condicional del salario, $p_{Y|X=j}(\cdot)$, un color para cada j nivel educativo



Esperanza condicional: $E(Y|X)$ es una variable aleatoria, que toma los siguientes valores:

j	5	12	15	18.50
$E(Y X = j)$	11.68	17.08	24.74	38.78

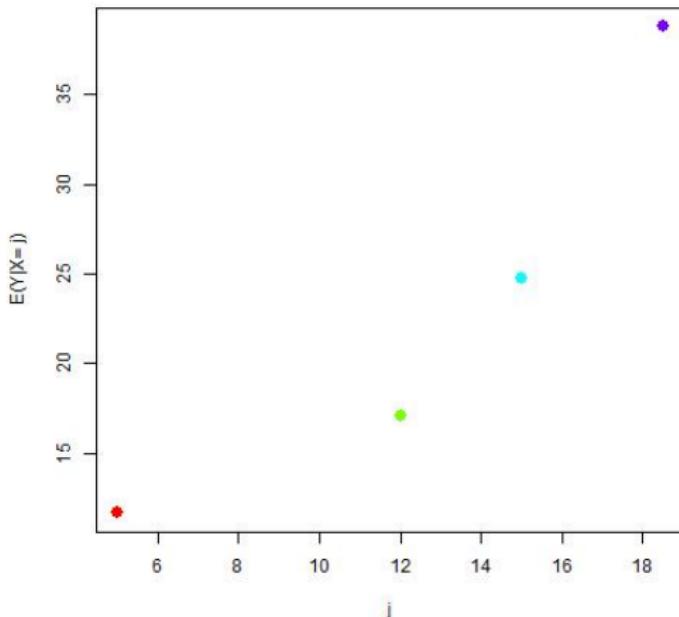


Gráfico de la esperanza condicional: $E(Y|X = j)$.
Conclusión:

- a mayor cantidad de años de estudio, mayor salario por hora, en promedio. 😊
- Además, se ve una relación exponencial entre salario y estudio.

Ejemplo 9.3: Distribución Normal Bivariada

Un vector aleatorio (X, Y) tiene distribución normal bivariada si su función de densidad conjunta es

$$f_{XY}(x, y) = \frac{1}{C} e^{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) \right]}.$$

con $C = 2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}$. La función de densidad depende de cinco

constantes:

$$\begin{aligned} -\infty < \mu_X < \infty &\quad -\infty < \mu_Y < \infty \\ \sigma_X > 0 &\quad \sigma_Y > 0 \quad -1 < \rho < 1 \end{aligned}$$

Las curvas de nivel son elipses. Las distribuciones marginales de X e Y son $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, respectivamente.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp \left\{ -\frac{1}{2\sigma_X^2} \left(\frac{x-\mu_X}{\sigma_X} \right)^2 \right\}$$

Entonces, la densidad condicional de Y dado X está dada por:

$$f_{Y|X=x}(y) = \frac{f_{XY}(x, y)}{f_X(x)}$$

Después de completar cuadrados, y otros cálculos, obtenemos que

$$f_{Y|X=x}(y) = \frac{1}{\sigma_Y \sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{1}{2} \frac{\left[y - \mu_Y - \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X)\right]^2}{\sigma_Y^2 (1 - \rho^2)}\right)$$

Esta es la densidad de una normal con media $\mu_Y + \rho(x - \mu_X)\sigma_Y/\sigma_X$ y varianza $\sigma_Y^2(1 - \rho^2)$. La distribución condicional de Y dado X es una normal univariada. Más aún, de esto se deduce la

esperanza condicional de Y dado X . Y también la varianza condicional :

$$E(Y|X) = \mu_Y + \rho(X - \mu_X)\sigma_Y/\sigma_X \quad \text{lineal}$$

$$V(Y|X) = \sigma_Y^2(1 - \rho^2) \quad \text{constante}$$

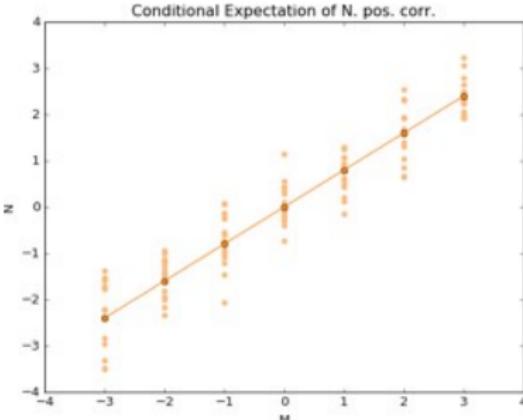
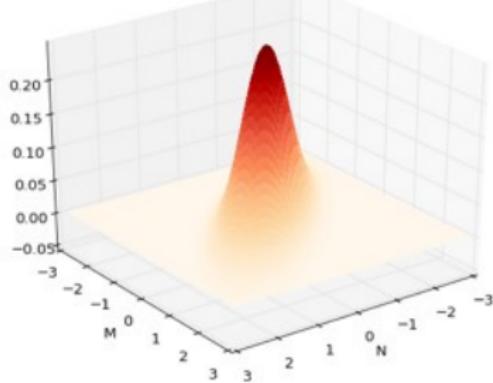
Visualización de la distribución normal bivariada

<http://jlcoto.github.io/visualizing-bivariate-normal>

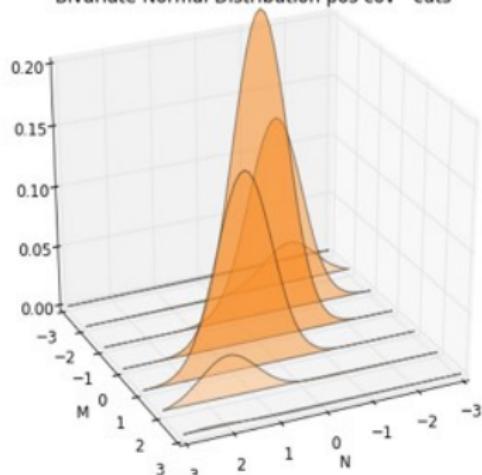
Los gráficos son

- 11 la función de densidad conjunta de un vector (M, N) con $\mu_M = \mu_N = 0$, $\sigma_M = \sigma_N = 1$ y $\rho(M, N) = 0,8$.
- 12 Observaciones de ambas variables aleatorias, tomadas a niveles fijos de $M = m_0$, con m_0 entero, $m_0 = -3, -2, -1, \dots, 3$ y la $E[N|M = m_0]$ superpuesta.
- 21 gráfico de cortes (secciones) de la función de densidad conjunta en distintos valores de $M = m$. Observar las distintas alturas de los cortes (que reflejan zonas de distinta probabilidad conjunta), de hecho ninguna es una función de densidad (no integran 1), graficamos $f_{MN}(m_0, n)$ con $n \in [-3, 3]$, m_0 fijo, $m_0 \in \{-3, -2, -1, 0, 1, 2, 3\}$.
- 22 gráfico de cortes (secciones) de la función de densidad condicional de $N | M$ en distintos valores de $M = m_0$. Observemos que todos tienen la misma altura, la misma dispersión, sus medias están linealmente alineadas. Son las mismas secciones de la densidad conjunta, reescaladas para integrar uno.

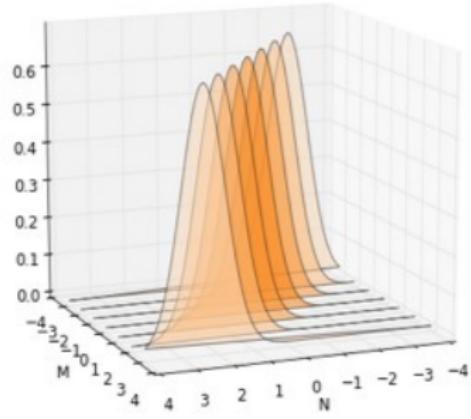
Bivariate Normal Distribution - covariance 0.8



Bivariate Normal Distribution pos cov - cuts



Conditional distributions at cuts - pos. corr.



Ejercicio 9.1 (Dos vectores uniformes (cont.))

Volvamos al Ejemplo 9.1 de la página 18. Sea $X \sim \mathcal{U}(0, 1)$,

$Y | X = x \sim \mathcal{U}(x, x + 1)$.

- a) Probar que $(X, Y) \sim \mathcal{U}(R)$ con R como en la Figura 2 a), de la página 19.
- b) Usando las distribución marginal de X y la condicional de $Y | X$ hallar $E(X)$, $V(X)$, $E(Y | X)$, $V(Y | X)$ y $E(XY | X)$.
- c) Usando estas tres variables calculadas en el ítem anterior, hallar $E(Y)$, $V(Y)$, $E(XY)$ y $\text{cov}(X, Y)$. Verificar los valores de la Tabla de la página 20.
- d) Comprobar que en este caso el mejor predictor de Y basado en X (es decir, $E(Y | X)$) resulta ser el óptimo lineal.

Ejemplo 9.4: condicionar al mínimo

Sea $Y \sim \text{Exp}(1)$. Sea $X = \min\{Y, 3\}$.

- Postular un candidato razonable para $E(Y|X)$ basándose en argumentos intuitivos y luego probar ^{que} el candidato propuesto es efectivamente $E(Y|X)$.
- Lo mismo para $E(X|Y)$.

Observemos que (X, Y) no tiene densidad conjunta en este caso. Es más, Y es una v.a. continua y X es una v.a. mixta (ni continua, ni discreta).

Si $X=t$ con $t < 3$, entonces mi mejor predicción del valor de Y es, precisamente, t .

Si $X=3$, entonces sabemos que $Y > 3$ y nada más sobre Y . Buscamos un valor $a > 3$ como valor para $E(Y|X=3)$. Luego, postulamos:

$$E(Y|X=t) = \begin{cases} t & \text{si } 0 \leq t < 3 \\ a & \text{si } t = 3 \end{cases}$$

¿Cuánto valdrá a ?

$$\text{Sea } g^*(t) = t I_{[0,3)}(t) + a I_{\{3\}}(t)$$

Elijamos a para que $E(Y|X) = g^*(X)$

- Claramente $g^*(X)$ es función de X .
- Sea $h(x)$. Queremos probar que

$$E([Y - g^*(X)] h(X)) = 0 \quad (\text{A})$$

Calculemos

$$\begin{aligned} E(g^*(X)h(X)) &= E\left(X I_{[0,3)}(X) h(X)\right) + a E\left(I_{\{3\}}(X) h(X)\right) \\ &= E(Y I_{[0,3)}(Y) h(X)) + a E(I_{[3,+\infty)}(Y) h(3)) \end{aligned} \quad (\text{B})$$

$$X I_{[0,3)}(X) = Y I_{[0,3)}(Y) \quad \text{y}$$

$$I_{\{3\}}(X) h(X) = h(3) I_{\{3\}}(X) = h(3) I_{[3,+\infty)}(Y).$$

y también

$$E(Y h(x)) = E\left(Y I_{[0,3)}(y) h(x)\right) + E\left(Y I_{[3,+\infty)}(y) h(x)\right)$$
$$= E\left(Y I_{[0,3)}(y) h(x)\right) + h(3) E\left(Y I_{[3,+\infty)}(y)\right) \quad (c)$$

$$Y I_{[3,+\infty)}(y) h(x) = Y I_{[3,+\infty)}(y) h(3).$$

$$(A) \Leftrightarrow (B) = (C) \Leftrightarrow$$

$$h(3) \alpha E\left(I_{[3,+\infty)}(y)\right) = h(3) E\left(Y I_{[3,+\infty)}(y)\right)$$

$\cancel{+ h.}$

$$a h(3) P(Y \geq 3) = E(Y I_{[3,+\infty)}(Y) \cdot h(3))$$

$$\Leftrightarrow a = \frac{E(Y I_{[3,+\infty)}(Y))}{P(Y \geq 3)} \quad \begin{matrix} \text{ambas probabilidades} \\ \text{se calculan con} \\ \text{la densidad de } Y. \end{matrix}$$

$$P(Y \geq 3) = \int_3^{+\infty} e^{-y} dy = -e^{-y} \Big|_3^{+\infty} = e^{-3}$$

$$\begin{aligned} E(Y I_{[3,+\infty)}(Y)) &= \int_3^{+\infty} y e^{-y} dy = (-e^{-y})(y+1) \Big|_3^{+\infty} \\ &= e^{-3} 4. \end{aligned}$$

$$\therefore a = \frac{4e^{-3}}{e^{-3}} = 4. \quad \begin{matrix} \leftarrow \text{Es la esperanza de } Y \text{ cuando} \\ \text{se restringe a que } Y \geq 3. \end{matrix}$$

Finalmente

$$E(Y|X) = \begin{cases} X & \text{si } 0 \leq X < 3 \\ 4 & \text{si } X = 3. \end{cases}$$

b) $E(X|Y) = E(\min\{Y, 3\} | Y) = \min\{Y, 3\}$

por propiedades de
esperanza condicional

Ejemplo 9.5: mezcla de dos distribuciones

Sean $\gamma | x \sim N(160 + 15x, (5-x)^2)$.

$x \sim Be(0,52)$ Sea $\alpha = 0,52$

- Hallar f_y , $E(y)$, $V(y)$. Graficar f_y . Interpretar.
- Para $x \in \{0,1\}$ calcular $P_{X|Y=y}(x)$. Calcular esta probabilidad para $x=0$ e $y=150, 160, 165, E(y), 170, 175$
- $F_y(y) = P(Y \leq y) = E(I_{(-\infty, y]}(y))$
 $P(Y \leq y) = P(Y \leq y, X=0) + P(Y \leq y, X=1)$
Probab total pues X es discreta

$$P(Y \leq y) = P(Y \leq y, X=0) + P(Y \leq y, X=1)$$

Probab total

$$= p_X(0) \cdot P(Y \leq y | X=0) + p_X(1) P(Y \leq y | X=1)$$

$$= (1-\alpha) P(W_0 \leq y) + \alpha P(W_1 \leq y)$$

con $W_0 \sim N(160, 36)$

$W_1 \sim N(175, 25)$

Si llamamos

f_{W_0} y f_{W_1} a sus densidades

$$F_Y(y) = (1-\alpha) F_{W_0}(y) + \alpha F_{W_1}(y).$$

respectivas

$$\Rightarrow f_Y(y) = (1-\alpha) f_{W_0}(y) + \alpha f_{W_1}(y)$$

Tanto F_Y como f_Y son combinaciones convexas:

de F_{W_0} y F_{W_1} , la distribución, y de f_{W_0} y f_{W_1} , la dens.

$$\begin{aligned}
 \boxed{E(Y)} &= E\left(\underbrace{E(Y|X)}_{g(x)}\right) = g(0) \cdot p_X(0) + g(1) p_X(1) \\
 &= E(Y|X=0) p_X(0) + E(Y|X=1) p_X(1) \\
 &= 160 \cdot 0,48 + 175 \cdot 0,52 = \boxed{167,8}
 \end{aligned}$$

O, también:

$$\begin{aligned}
 E(Y) &= E(E(Y|X)) = E(160 + 15X) = 160 + 15 E(X) \\
 &= 160 + 15 \cdot 0,52 = 167,8.
 \end{aligned}$$

Podemos hallar $\text{Var}(Y)$ de dos formas distintas.

1º Usando que

$$\text{Var}(Y) = \text{Var}(\mathbb{E}(Y|X)) + \mathbb{E}(\text{Var}(Y|X)).$$

$$Y|X \sim N(160 + 15X, (6 - X)^2).$$

$$\begin{aligned}\text{Var}(\mathbb{E}(Y|X)) &= \text{Var}(160 + 15X) = \text{Var}(15X) = 15^2 \cdot \text{Var}(X) \\ &= 15^2 \cdot (0,52)(1 - 0,52) = 56,16\end{aligned}$$

$\uparrow X \sim \text{Be}(\alpha)$

$$\begin{aligned}\mathbb{E}(\text{Var}(Y|X)) &= \mathbb{E}((6 - X)^2) = \mathbb{E}(36 - 12X + X^2) \\ &= 36 - 12 \cdot \mathbb{E}(X) + \mathbb{E}(X^2) \underset{\uparrow}{=} 36 - 6,24 + 0,52 \\ &= 30,28\end{aligned}$$

$\boxed{\mathbb{E}(X) = \mathbb{E}(X^2) = 0,52}$

Luego $\text{Var}(Y) = 56,16 + 30,28 = 86,44$

2^a forma) Usando que la densidad es una comb. convexa.

$$E(Y^2) = \int_{-\infty}^{+\infty} y^2 f_Y(y) dy =$$

$$= \alpha \int_{-\infty}^{+\infty} y^2 f_{W_1}(y) dy + (1-\alpha) \int_{-\infty}^{+\infty} y^2 f_{W_0}(y) dy$$

$$= \alpha E(W_1^2) + (1-\alpha) E(W_0^2)$$

$$= \alpha [V(W_1) + E(W_1)^2] + (1-\alpha) [V(W_0) + E(W_0)^2]$$

$$= \alpha [25 + (175)^2] + (1-\alpha) [36 + (160)^2]$$

Finalmente

$$V(Y) = E(Y^2) - E(Y)^2 = 86,44$$

Observemos que $V(Y)$ es mucho mayor que la varianza de Y cuando conocemos el valor de X :

$$V(Y|X=0) = 36 \quad V(Y|X=1) = 25$$

Observamos que puede pensarse que este proceso corresponde a un experimento en 2 etapas.

→ La 1^a sorteó una moneda (X)

$$X=0$$

$$X=1$$

→ La 2^a elige una v.a. continua (Y) con distinta distribución de acuerdo al valor de X obtenido.

→ si $X=0 \rightarrow Y \sim N(160, 36)$

→ si $X=1 \rightarrow Y \sim N(175, 25)$.

Podemos pensar que estamos eligiendo una persona en una población, y luego medimos su altura.

→ 1º sorteamos su género

$\begin{cases} X=0 & \text{mujer} \\ X=1 & \text{varón} \end{cases}$

(52% hombres)

→ 2º condicional a que sabemos su género, sabemos que

→ La altura de los hombres (en cm.) es una v.a. $N(\mu = 175, \sigma^2 = 25)$.

→ La altura de las mujeres es una v.a. con distribución $N(\mu = 160, \sigma^2 = 36)$

Figura 5: Gráfico de la densidad de Y , f_Y

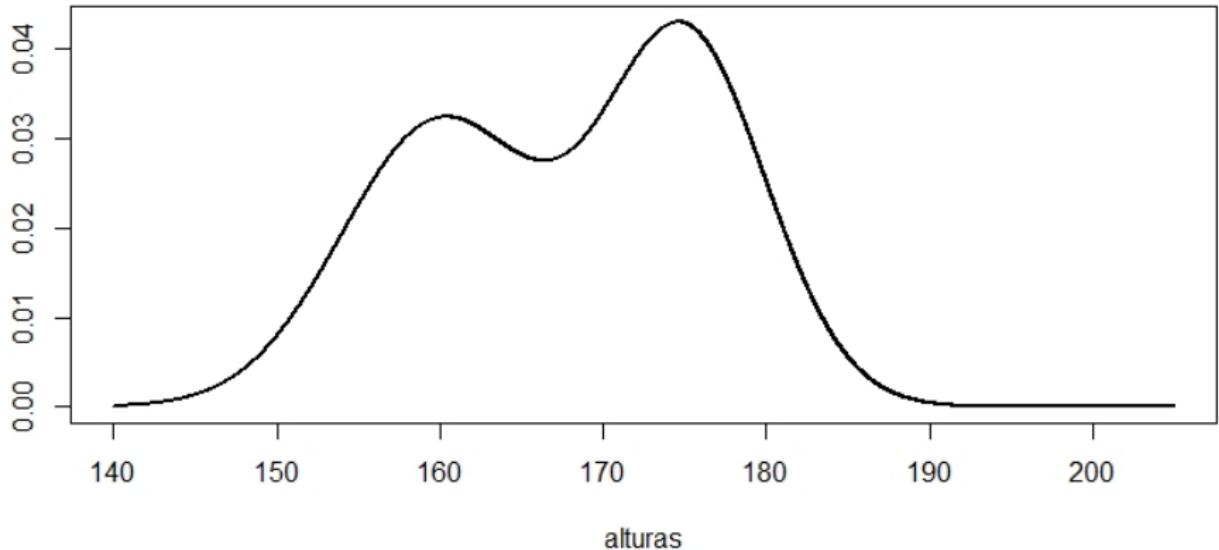
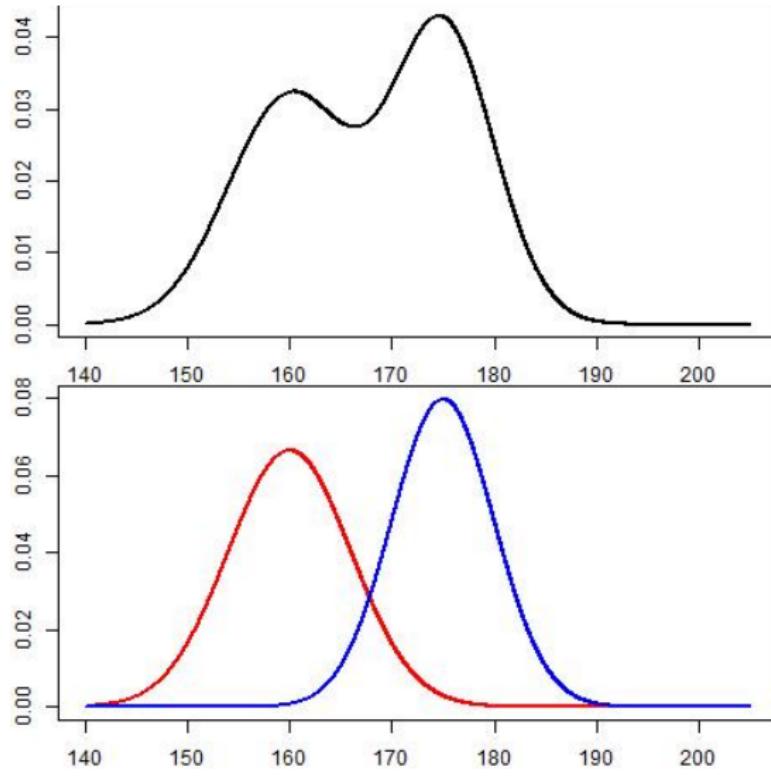


Figura 6: Gráfico de la densidad de Y , f_Y (mezcla) arriba y de las dos condicionales abajo, $f_{Y|X=0}$ y $f_{Y|X=1}$



El fenómeno de mezclar 2 o más poblaciones homogéneas produce una densidad con más de un máximo local (o moda), y aumenta las varianzas originales.

Da una F_y (resp. f_y) que es una comb convexa de funciones de distribución (resp. densidades). La esperanza es una combinación convexa de las originales.

La varianza no. Resulta mayor

b) Queremos $p_{x|y=y}(x) = E(I_{\{x\}}(x) | Y=y)$

Si valiera una versión general de Bayes tendríamos:

$$p_{x|y=y}(x) = \begin{cases} \frac{f_{Y|x=x}(y) p_x(x)}{f_y(y)} & \text{si } f_y(y) > 0 \\ 0 & \text{sino.} \end{cases}$$

dónde $f_y(y) = \sum_{x \in R_X} p_x(x) f_{Y|x=x}(y)$ ¿Será?

Probémoslo. Sea $g(y) = \begin{cases} \frac{f_{Y|x=x}(y) p_x(x)}{f_y(y)} & \text{si } f_y(y) > 0 \\ 0 & \text{si } f_y(y) = 0 \end{cases}$

La esperanza condicional de $I_{\{x\}}(x)$ dado Y será $g_z(Y)$: siempre que logremos probar que, $E(g_z(Y)) = E(I_{\{x\}}(x) h(Y))$:

$$E(I_{\{x\}}(x) h(Y)) = E(g_z(Y) h(Y)).$$

$$g(y) = \begin{cases} \frac{f_{Y|X=x}(y) p_x(x)}{f_y(y)} & \text{si } f_y(y) > 0 \\ 0 & \text{si } f_y(y) = 0 \end{cases}$$

$$\mathbb{E}(g(Y) h(Y)) = \int_{-\infty}^{+\infty} g(y) h(y) f_y(y) dy$$

Yes va cont

$$= \int_{-\infty}^{+\infty} \frac{f_{Y|X=x}(y) p_x(x)}{f_y(y)} h(y) f_y(y) dy.$$

reemplazamos $g(y)$ por la propuesta

$$= \int_{-\infty}^{+\infty} f_{Y|X=x}(y) p_x(x) h(y) dy$$

Por otro lado:

$$E(I_{\{x\}}(x) h(y)) = E(\underbrace{E(\varphi(x, y) | x)}_{g^*(x)})$$

$\varphi(x, y)$

$$= \sum_{k \in R_x} g^*(k) \cdot p_x(k)$$

$$= \sum_{k \in R_x} p_x(k) \int_{-\infty}^{+\infty} \varphi(k, y) f_{y|x=k}(y) dy$$

$$= \int_{-\infty}^{+\infty} \sum_{k \in R_x} p_x(k) \underbrace{I_{\{x\}}(k) h(y)}_{=1 \text{ solo si } k=x} f_{y|x=k}(y) dy$$

reemplazamos

$$\varphi(k, y)$$

$$= \int_{-\infty}^{+\infty} p_x(x) h(y) f_{y|x=x}(y) dy$$

Cuadro 2: Calculamos la probabilidad de que $X = 0$ condicional a $Y = y$, es decir, $p_{X|Y=y}(0)$ para distintos valores de y . Se puede interpretar como la probabilidad de que la persona elegida sea mujer, sabiendo que mide y centímetros. Vemos que esa probabilidad baja a medida que y crece. ¿Cuál sería la probabilidad de $X = 0$ sin tener ninguna información sobre Y ?

y	$p_{X Y=y}(0)$
150	0.99998
160	0.98576
165	0.80066
167.8	0.48237
170	0.24026
175	0.03269
y	$\frac{f_{Y X=0}(y)p_X(0)}{f_Y(y)}$

Condicionando a Vectores

Hasta el momento hemos trabajado con una sola variable predictora. ¿Cómo hacer si la información disponible para predecir es un vector? Por ejemplo, en los datos de salario en EE.UU. podríamos disponer de otras variables aleatorias medidas en la base de datos:

- X_1 = género del trabajador o trabajadora
- X_2 = educación (medida en años)
- X_3 = indicadora de estar afiliado a un sindicato
- X_4 = antigüedad en el actual empleo
- X_5 = edad
- Y_1 = salario por hora
- Y_2 = gasto total mensual del hogar

En este caso, tenemos $\tilde{\mathbf{X}} = (X_1, X_2, X_3, X_4, X_5) \in \mathbb{R}^5$ e $\tilde{\mathbf{Y}} = (Y_1, Y_2) \in \mathbb{R}^2$ dos vectores aleatorios. Nos interesa usar el vector $\tilde{\mathbf{X}}$ para predecir funciones del vector $\tilde{\mathbf{Y}}$, como por ejemplo

$$E[Y_1 | \tilde{\mathbf{X}}], E[Y_2 | \tilde{\mathbf{X}}], E[Y_1 \times 40 \times 4 - Y_2 | \tilde{\mathbf{X}}]$$

o más generalmente,

$$E[h(Y_1, Y_2) | \tilde{\mathbf{X}}].$$

Sea $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ un nuevo vector aleatorio cuyas coordenadas son vectores aleatorios en los espacios \mathbb{R}^k y \mathbb{R}^j respectivamente:

$$\tilde{\mathbf{X}} \in \mathbb{R}^k, \quad \tilde{\mathbf{Y}} \in \mathbb{R}^j,$$

como podemos definir

$$E[g(Y_1, Y_2, \dots, Y_j) | X_1, X_2, \dots, X_k] ?$$

Claramente, estamos queriendo extender la noción de

$$E[I(Y) | X]$$

al caso en el que tanto X como Y pueden ser vectores. Cuando el vector conjunto $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ es discreto o continuo, definimos la distribución condicional del vector $\tilde{\mathbf{Y}}$ condicionado al vector $\tilde{\mathbf{X}}$ y, una vez definida la distribución condicional, calcularemos esperanzas según la distribución condicional.

Definición 9.9

Sea $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ un vector aleatorio en \mathbb{R}^{k+j} , siendo que $\tilde{\mathbf{X}} \in \mathbb{R}^k$, $\tilde{\mathbf{Y}} \in \mathbb{R}^j$. Para cada \mathbf{x} con $p_{\tilde{\mathbf{X}}}(\mathbf{x}) > 0$ o $f_{\tilde{\mathbf{X}}}(\mathbf{x}) > 0$, definimos la distribución condicional de $\tilde{\mathbf{Y}}$ dado que $\tilde{\mathbf{X}} = \mathbf{x}$ mediante

$$p_{\tilde{\mathbf{Y}}|\tilde{\mathbf{X}}=\mathbf{x}}(\mathbf{y}) = \frac{p_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}(\mathbf{x}, \mathbf{y})}{p_{\tilde{\mathbf{X}}}(\mathbf{x})} \text{ cuando el vector sea discreto,}$$

$$f_{\tilde{\mathbf{Y}}|\tilde{\mathbf{X}}=\mathbf{x}}(\mathbf{y}) = \frac{f_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}(\mathbf{x}, \mathbf{y})}{f_{\tilde{\mathbf{X}}}(\mathbf{x})} \text{ cuando el vector sea continuo.}$$

En esta definición $p_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$ denota la probabilidad puntual del vector $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$, mientras que $p_{\tilde{\mathbf{X}}}$ denota la probabilidad marginal asociada al vector $\tilde{\mathbf{X}}$, en el caso discreto, mientras que $f_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$ denota la densidad conjunta del vector $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$, mientras que $f_{\tilde{\mathbf{X}}}$ denota la densidad marginal asociada al vector $\tilde{\mathbf{X}}$.

Tras haber definido la distribución condicional, tenemos que la esperanza condicional está dada por la siguiente fórmula:

Lema 9.10

El mejor predictor para $g(\tilde{\mathbf{Y}})$ basado en $\tilde{\mathbf{X}}$ está dado por la esperanza de la función $g(\tilde{\mathbf{Y}})$ respecto de la distribución de $\tilde{\mathbf{Y}}$ condicionada a $\tilde{\mathbf{X}}$. Es decir, es una variable aleatoria, que denotaremos por $E[g(\tilde{\mathbf{Y}}) | \tilde{\mathbf{X}}]$ que cuando el vector aleatorio $\tilde{\mathbf{X}}$ toma los valores $(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$, vale

$$E[g(Y_1, Y_2, \dots, Y_j) | X_1 = x_1, X_2 = x_2, \dots, X_k = x_k] =$$

$$= \begin{cases} \sum_{\mathbf{y}} g(\mathbf{y}) p_{\tilde{\mathbf{Y}}|\tilde{\mathbf{X}}=\mathbf{x}}(\mathbf{y}) & \text{si el vector es discreto} \\ \int_{-\infty}^{\infty} g(\mathbf{y}) f_{\tilde{\mathbf{Y}}|\tilde{\mathbf{X}}=\mathbf{x}}(\mathbf{y}) d\mathbf{y} & \text{si el vector es continuo,} \end{cases}$$

donde $\mathbf{y} = (y_1, y_2, \dots, y_j)$ y $\mathbf{x} = (x_1, x_2, \dots, x_k)$

Más generalmente, si conocemos la distribución de $\tilde{\mathbf{Y}} | \tilde{\mathbf{X}} = \mathbf{x}$, definimos $E[g(\tilde{\mathbf{Y}}) | \tilde{\mathbf{X}} = \mathbf{x}]$ como la esperanza de g según dicha distribución.

Además, valen las mismas propiedades del Teorema 9.5 con vectores en lugar de variables aleatorias, que recuperamos en el siguiente Teorema.

También vale la versión con vectores de la condición de mejor predictor de la esperanza condicional, enunciada en el Teorema 9.4

Teorema 9.11 (Propiedades de esperanza condicional para vectores)

Propiedades de la esperanza condicional:

- ① $E\left[E[g(\tilde{\mathbf{Y}}) | \tilde{\mathbf{X}}]\right] = E[g(\tilde{\mathbf{Y}})]$
- ② $E[g_1(Y_1) + g_2(Y_2) | \tilde{\mathbf{X}}] = E[g_1(Y_1) | \tilde{\mathbf{X}}] + E[g_2(Y_2) | \tilde{\mathbf{X}}]$
- ③ $E[cg(\tilde{\mathbf{Y}}) | \tilde{\mathbf{X}}] = cE[g(\tilde{\mathbf{Y}}) | \tilde{\mathbf{X}}], c \in \mathbb{R}$.
- ④ $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \text{ independientes, entonces } E[g(\tilde{\mathbf{Y}}) | \tilde{\mathbf{X}}] = E[g(\tilde{\mathbf{Y}})]$
- ⑤ $E[g(\tilde{\mathbf{X}}) | \tilde{\mathbf{X}}] = g(\tilde{\mathbf{X}})$
- ⑥ $E[r(\tilde{\mathbf{X}})g(\tilde{\mathbf{Y}}) | \tilde{\mathbf{X}}] = r(\tilde{\mathbf{X}})E[g(\tilde{\mathbf{Y}}) | \tilde{\mathbf{X}}]$
- ⑦ Propiedad de torres:

$$\begin{aligned} & E\left[E[h(\tilde{\mathbf{Y}}) | X_1, X_2, \dots, X_g, \dots, X_k]\right] | X_1, X_2, \dots, X_g \\ &= E[h(\tilde{\mathbf{Y}}) | X_1, X_2, \dots, X_g] \end{aligned}$$

para todo $g \leq k$.

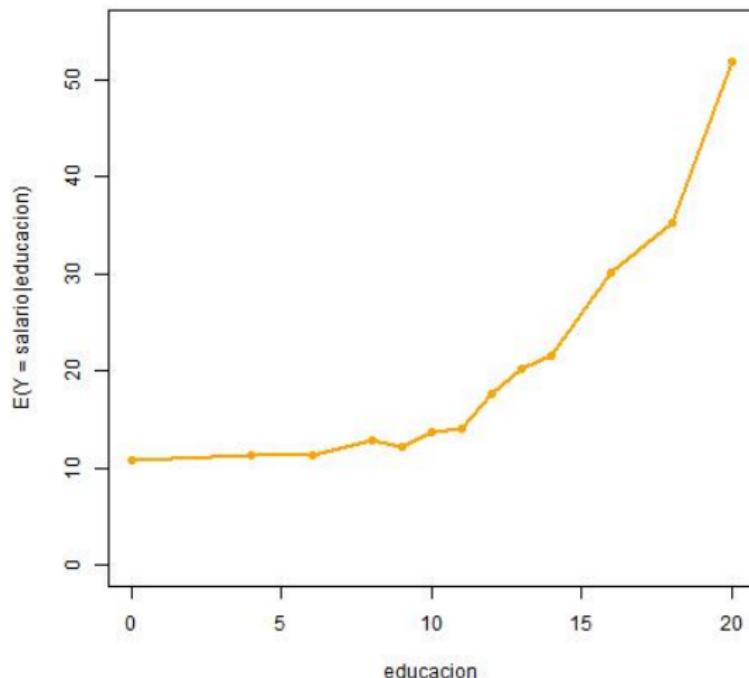
Ejemplo 9.2: Salarios EE.UU. (segunda parte)

Volvamos al ejemplo de los salarios, miremos a la variable salario por hora como continua. La variable educación toma 13 valores distintos. Un gráfico parecido al de la Figura 4 sería imposible de mirar, uno como la Figura 3 sería muy difícil.

Sin embargo, podemos hacer un gráfico de la esperanza del salario por hora condicional al nivel de educación (salario medio por nivel educativo).

Recordemos que estamos pensando en los 50.742 trabajadores de la base de datos como si fuera el total de la población.

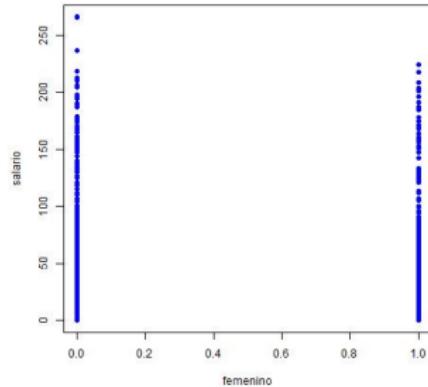
Figura 7: Salario (por hora) medio por nivel educativo, i.e. $E(Y | X)$, donde Y = salario (por hora), y X = educación. Vemos que a mayor cantidad de años de educación recibida, mayor salario esperado.



La base de datos tiene otras variables medidas en cada trabajador. Por ejemplo, el género. Esta variable se denomina **femenino** y vale 1 si el trabajador es una mujer. Entonces surge la pregunta: **¿cobrarán más los hombres o las mujeres?**

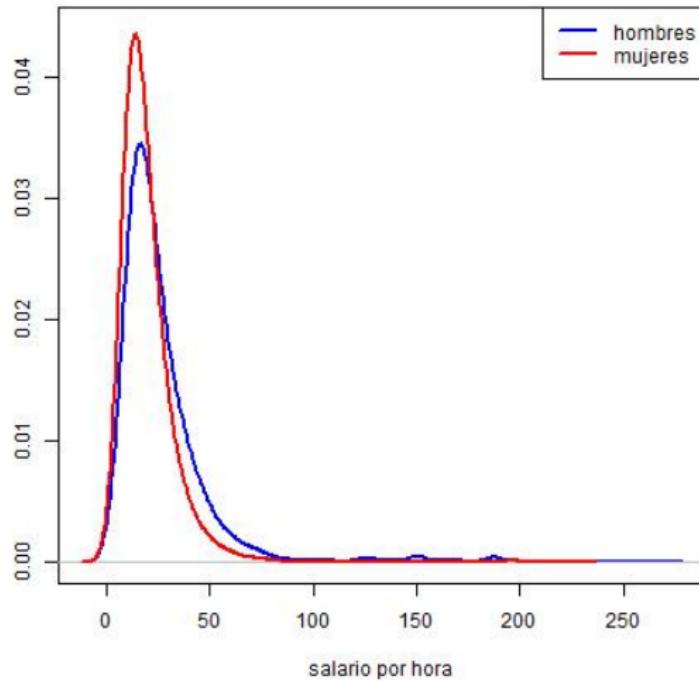
Figura 8: Salario (por hora) por género
(femenino = 0 es varón).

Podemos graficar cada observación, separada por género para tratar de contestar a la pregunta.



No se ve nada

Figura 9: Histograma suavizado del salario (por hora), separado por género.



Las densidades parecen diferentes, quizá la curva de las mujeres más corrida hacia el cero. ¿Será que en promedio el salario de las mujeres es menor al de los varones?

Para contestarla, calculamos $E(Y | \text{femenino})$

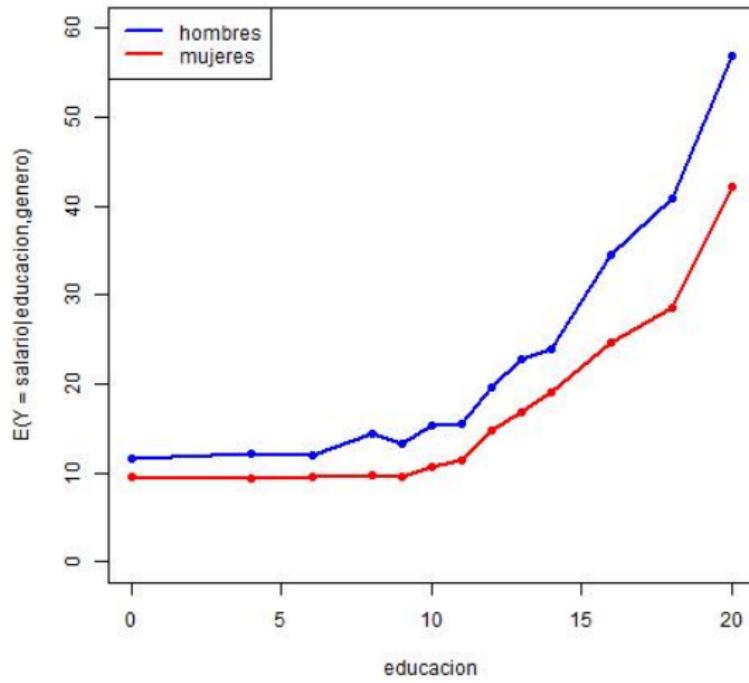
femenino	$E(Y \text{femenino})$
0	26.80
1	20.00

¡Hay mucha diferencia en el salario medio!

¿Cobrarán menos las mujeres por ser mujeres? ¿O será, por ejemplo, que las mujeres recibieron además menos años de educación?

Para ver si esto sucede, o no, podemos hallar la esperanza del salario por hora, condicional al nivel educativo y al género, o sea, calcular
 $E(\text{salario} | (\text{género}, \text{educación}))$

Figura 10: Salario medio (por hora) por género y por educación, $E(\text{salario} \mid (\text{género}, \text{educación}))$.



Conclusión: ¡¡Las mujeres cobran menos!!

Ejercicio 9.2 (Bayes 1)

Sea (X, Y) un vector aleatorio con X una variable aleatoria discreta, con rango finito, $R_X = \{x_i : 1 \leq i \leq n\}$. Además, para cada valor $x \in R_X$ tenemos que la distribución de $Y | X = x$ es continua con función de densidad condicional dada por $f_{Y|X=x}(y)$.

- (i) Probar que Y es una variable aleatoria continua y calcular su densidad, f_Y . Sug: Hallar $F_Y(y)$, condicionando adecuadamente.
- (ii) Probar **por definición de esperanza condicional** que para todo $x \in R_X$ tenemos

$$p_{X|Y=y}(x) = \begin{cases} \frac{f_{Y|X=x}(y) \cdot p_X(x)}{f_Y(y)} & \text{si } f_Y(y) \neq 0 \\ 0 & \text{si } f_Y(y) = 0 \end{cases} \quad (5)$$

¿Qué significa probar (ii)? Significa que si llamamos $g_x^*(y)$ al lado derecho de la igualdad (5), entonces vale que $g_x^*(Y) = P(X = x | Y) = E(I_{\{x\}}(X) | Y)$ en el sentido de predictor óptimo.

Ejercicio 9.3 (Bayes 2)

Sea (X, Y) un vector aleatorio con X una variable aleatoria discreta, con rango finito, $R_X = \{x_i : 1 \leq i \leq n\}$ e Y una variable continua con función de densidad f_Y . Tenemos que la distribución de $X | Y = y$ es discreta con función de probabilidad puntual condicional dada por $p_{X|Y=y}(x)$.

- (i) Probar que para todo $x \in R_X$ se tiene que

$$p_X(x) = \int_{-\infty}^{+\infty} p_{X|Y=y}(x) f_Y(y) dy.$$

- (ii) Probar *por definición de esperanza condicional* que para todo $x \in R_X$ tenemos

$$f_{Y|X=x}(y) = \frac{p_{X|Y=y}(x) \cdot f_Y(y)}{p_X(x)} \quad (6)$$

¿Qué significa probar (ii)? Significa que la expresión dada por (6), cumple las propiedades de una densidad condicional, es decir, que para todo $a \in \mathbb{R}$ vale que $P(Y \in (-\infty, a] | X = x) = \int_{-\infty}^a f_{Y|X=x}(y) dy := g_a^*(x)$. Ergo, si definimos $g_a^*(x)$ como la expresión anterior donde $f_{Y|X=x}(y)$ está definida en (6), entonces para cada $a \in \mathbb{R}$, $g_a^*(X)$ cumple la definición de predictor óptimo: $E(I_{(-\infty, a]}(Y) | X) = g_a^*(X)$

Observación 9.1

- Los dos ejercicios anteriores son válidos aún sin la hipótesis de que el rango de la variable discreta sea finito. La agregamos simplemente para facilitar las cuentas, ya que para probarlo en el caso general se usan resultados de teoría de la medida.
- Observemos que en el ejercicio 9.3 se pide probar la fórmula (6) como densidad condicional. ¿Por qué debería ser la distribución de $Y | X = x$ (absolutamente) continua, para cada $x \in R_X$? Una forma de verlo: vimos que Y es una variable (abs) continua si y sólo $P(Y \in E) = 0$ para todo conjunto E de medida cero. Como el evento $\{X = x\}$ tiene probabilidad positiva, entonces

$$P(Y \in E | X = x) = \frac{P(\{Y \in E\} \cap \{X = x\})}{P(X = x)} \leq \frac{P(Y \in E)}{P(X = x)} = 0,$$

entonces la distribución de $Y | X = x$ es absolutamente continua para todo $x \in R_X$.

Observación 9.2

Tanto al resolver los ejercicios de la práctica, como los del parcial, se espera que deduzcan las distribuciones de la forma en que lo hicimos, por ejemplo, en el Ejemplo 9.5, ya que es una forma de practicar y afianzar las propiedades de esperanza y distribución condicional, y no que apelen a los resultados de los Ejercicios 9.2 y 9.3.

Ejercicio 9.4

Sean X e Y variables aleatorias tales que X es discreta, $Y \sim \beta(a, b)$ y $X | Y \sim Bi(n, X)$.

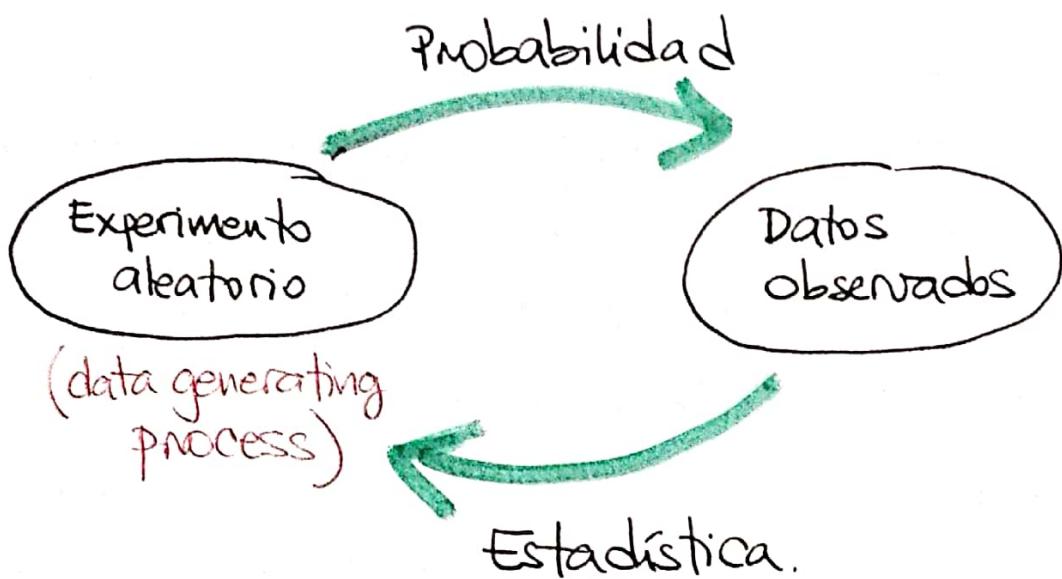
- (a) Hallar la distribución de Y y calcular, usando esperanza condicional, la $E(Y)$ y $V(Y)$.
- (b) Si a y b se eligen de modo tal que $X \sim U(0, 1)$, hallar la distribución de Y .

Ejercicio 9.5

Sean X e Y variables aleatorias tales que $X \sim Be\left(\frac{1}{3}\right)$ e $Y | X \sim U[1 - X, 2 + X]$.

- (a) Hallar F_Y y f_Y .
- (b) Para $x \in \{0, 1\}$ calcular $p_{Y|X=x}(y)$.

10. Estadística



La Probabilidad (de la que nos ocupamos hasta ahora) es el lenguaje formal de la incertezza.

El problema clave que estudia la probabilidad es: **dado un experimento aleatorio, cuáles son las propiedades de las observaciones que surgen de él?**

La Estadística (como el data mining y el machine learning) se ocupan de un problema en algún sentido inverso:

dados los datos observados (outcomes), ¿qué podemos decir del proceso o experimento aleatorio que los genera?

Miremos un par de ejemplos.

Ejemplo 1 (Control de calidad) Un importador de naranjas recibe un cargamento de $N = 10.000$ naranjas. Quiere saber cuántas de estas se pudrieron.

Para averiguarlo, toma una muestra de $n = 50$ naranjas. Un número aleatorio de ellas, x , están podridas. ¿Qué puede concluir sobre r , el número verdadero de naranjas podridas?

→ Hay 3 formas ^{clásicas} de contestar esta pregunta.

1^a) Con un **estimador puntual**: o sea, haciendo una cuenta con x y proponiendo un posible valor para r

2^a) Con una **estimación por intervalo** (o estimación más error): no sólo daremos una cuenta $R(x)$ que approxime a r si no que daremos un intervalo de extremos aleatorios tal que la probab de que ese intervalo contenga al verdadero r sea alta (pongamos 0,95).

3^a) Tomando una decisión entre 2 posibilidades
 ↳ comprar } el lote, mirando el valor x
 ↳ no comprar

Tomando en cuenta los errores posibles. Esto se llama **test estadístico**

Comenzamos por el 1º Estimación puntual.

¿Qué cuenta hacemos con x (cantidad de naranjas podridas entre las 50 revisadas) para estimar a Γ (cantidad verdadera de naranjas podridas en el lote)?



Podemos dar una respuesta intuitiva. Podemos argumentar que la proporción de naranjas podridas en la muestra es muy parecida a la proporción de naranjas podridas en la población. Osea:

$$\frac{x}{n} \approx \frac{\Gamma}{N}$$

De donde podemos proponer que aproximadamente

$$\Gamma \approx N \cdot \frac{x}{n}$$

y la cuenta será: $R(x) = \frac{Nx}{n}$

(Mejor dicho, el entero más cercano)

$$R(x) = \left[\frac{Nx}{n} \right]$$

es el estimador natural de Γ .

$R(x)$ será el ESTIMADOR

Obviamente, $R(x)$ es aleatorio.

Si el importador elige una segunda muestra,⁴ obtendrá un resultado diferente x' y por lo tanto, una estimación diferente $R(x')$.

Ejemplo 2: (Planta nuclear) En una planta nuclear se quiere monitorear la fragilidad de una tubería de enfriamiento. Se toman entonces n mediciones aleatorias x_1, \dots, x_n .

Cada medición puede pensarse como una suma de pequeñas perturbaciones, podemos asumir que cada medición $x_i \sim N(\mu, \sigma^2)$. Asumamos σ^2 conocida (depende de la precisión del método de medición, que podemos asumir pequeña). La esperanza de cada x_i , $\mu = E(x_i)$ con

x_i = iésima medición de la fragilidad de la tubería

μ es desconocida, y es la fragilidad real o verdadera. Queremos estimar a μ a partir de los valores de x_1, \dots, x_n observados.

Un estimador posible es usar el promedio:

$$\frac{1}{n} \sum_{i=1}^n x_i$$

Nos inspiramos en la LGN para proponerlo.

¿Cuál es la estructura detrás de los ejemplos? ⁵

En ambos casos hay un **experimento**, que produce **resultados numéricos** aleatorios. (datos observables)

↳ $x \in \{0, 1, \dots, n\}$ en el ejemplo 1.
↳ \mathbb{R}^n , en el ejemplo 2. $\xrightarrow{\text{Lo llamamos } x}$

La **distribución** en X que describe a las observaciones es desconocida: es lo que debemos identificar a través de los valores observados.

No consideramos una sola distribución de probabilidad en X sino muchas.

• En el ejemplo 1: $X \sim \text{Hipergeométrica}(n, r, N)$

N = total poblacional
 r = buenos en la población.
 n = cantidad de extraídos ($\frac{\text{tamaño de la muestra}}{\text{elementos}}$)

• En el ejemplo 2: $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ con

σ^2 conocida (precisión de las obs.).

μ desconocida

Definición: Un modelo estadístico es un conjunto de distribuciones (o funciones de densidad o funciones de probabilidad puntual).

Si dicho conjunto está parametrizado por un número finito de parámetros, se lo denomina un modelo paramétrico $\{F_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$.

A θ se lo denomina parámetro, y es una cantidad desconocida.

En el ejemplo 1: $\{J_t(n, r, N) : r \in \mathbb{N}\}$. (un parámetro)

En el ejemplo 2: $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}_{>0}, \sigma > 0\}$. (dos parámetros)

A Θ se lo llama espacio de parámetros. (tiene que contener al menos 2 valores distintos). ↑ Titón

Como en un modelo estadístico hay muchas distribuciones, debemos indicar con cual estamos trabajando al tomar esperanzas y varianzas.

Se suele escribir E_θ o V_θ .

La primera tarea importante es elegir el modelo adecuado.

Definición: El proceso de observar un experimento aleatorio que produce un resultado aleatorio será descrito a través de una variable aleatoria que, típicamente, denotaremos por X_i (θX). en su iésima repetición. A la colección X_1, \dots, X_n de raiid se las denomina una muestra aleatoria. En contraste, a la realiza-

ción de X o X_i , es decir, al valor específico x o x_i ⁷ observado se lo llamará resultado, dato o valor observado. (Notar mayúsculas para va. y minúsculas para valores observados).

Definición: Una función boreiana $g: \mathbb{R}^n \rightarrow \mathbb{R}$ se denomina un estadístico.

$T = g(X_1, \dots, X_n)$ es un estadístico.

Si el objetivo de la cuenta propuesta por T con la muestra es estimar un parámetro θ , a T se lo llama un estimador de θ y se lo nota $\hat{\theta}$ o también $\hat{\theta}_n$ o también $\hat{\theta}_n(X_1, \dots, X_n)$.

O sea: $T = g(X_1, \dots, X_n) = \hat{\theta}_n = \hat{\theta} = \hat{\theta}_n(X_1, \dots, X_n)$.

 mmm... pero entonces
¿un estadístico es un nombre
nuevo para una variable aleatoria?

Bueno, no cualquier v.a., una v.a.
que surge al hacer una cuenta
(o sea, aplicarle una función) a una
muestra aleatoria.

La figura que sigue, del Libro: Georgii, H.O. Stochastics: Introduction to Probability and Statistics, ilustra el proceso de estimación puntual

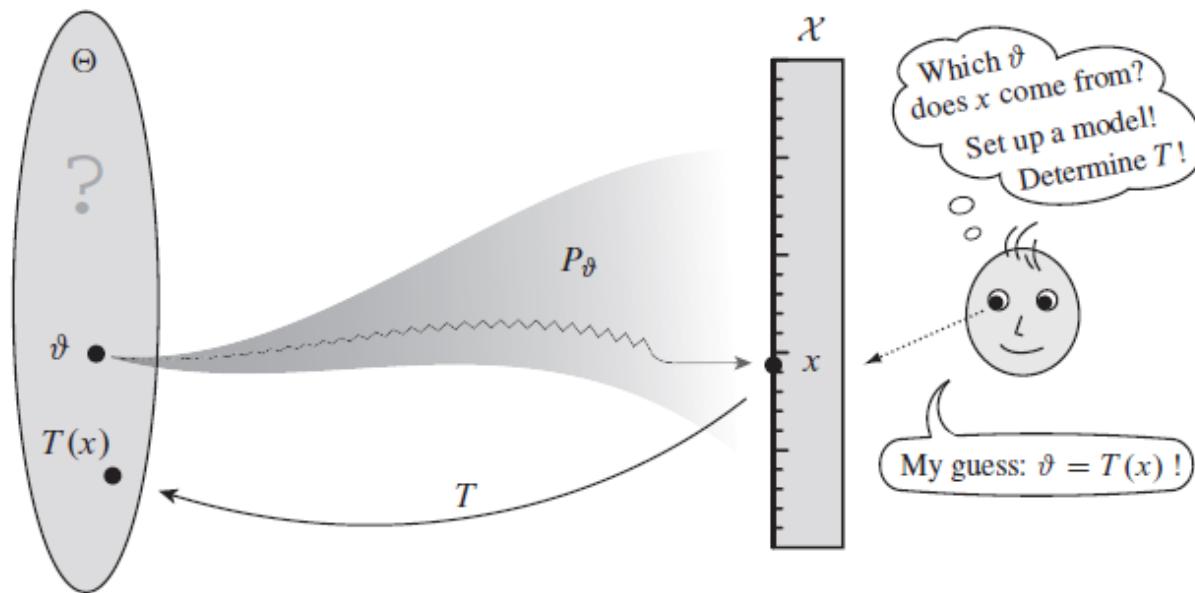


Figure 7.1. The principle of estimation: An unknown quantity ϑ is to be determined; the possible ϑ form a set Θ . The true ϑ determines a probability measure P_ϑ , which governs the random experiment and which leads to the observed value x in the concrete case. To infer ϑ , the statistician first has to analyse the probability measure P_ϑ for each ϑ – this is part of the set-up of the model. Then he must use the detected structure of all P_ϑ to construct a mapping $T : \mathcal{X} \rightarrow \Theta$ in such a way that the value $T(x)$ is typically close to the true ϑ – no matter which ϑ this happens to be in the end.

Ejemplo 3: (Adivinando el rango de ^{un generador de} números aleat.)

En un programa de entretenimientos se presenta una máquina que genera números aleatorios en el intervalo $[0, \theta]$, donde el anfitrión fijó el θ (pero nadie más lo sabe). Dos participantes pueden usar la máquina 10 veces ($n=10$) y luego tienen que adivinar el valor de θ . El que da el valor más cercano gana.

¿Cuál es el modelo estadístico en este caso? $\{U(0, \theta) : \theta \in \Theta = (0, +\infty)\}$

La muestra aleatoria consistirá en

$$x_1, \dots, x_n \sim U(0, \theta) \text{ vaid.}$$

entre los cuales está el verdadero θ_0 que da lugar a los datos.

Jugador A:

Este jugador se acuerda de la LGN

$$\text{Como } E_{\theta}(x_i) = \frac{\theta}{2} \text{ si } x_i \sim U(0, \theta).$$

Dice lo siguiente:

$$\bar{x}_n \xrightarrow{P} E(x_i) = \frac{\theta}{2} \text{ si } x_i \sim U(0, \theta)$$

$$\Rightarrow 2\bar{x}_n \xrightarrow{P} \theta \text{ si } x_i \sim U(0, \theta)$$

El jugador A elige $A_n = 2\bar{x}_n$ como estimador y cruza los dedos deseando que $n=10$ sea suf. gde

Jugador B:

Este jugador dice: aunque $X_{(n)} = \max\{X_1, \dots, X_n\}$ siempre será menor a θ , estará bastante cerca si n es grande. De hecho, tenemos: ($0 < t < \theta$)

$$F_{X_{(n)}}(t) = P(X_{(n)} \leq t) = P(X_1 \leq t, \dots, X_n \leq t) = \\ \underset{x_i \sim U(0, \theta), 1 \leq i \leq n}{\uparrow} = \prod_{i=1}^n P(X_i \leq t) = [F_{X_1}(t)]^n$$

Si $X_i \sim U(0, \theta)$

$$F_{X_i}(t) = \begin{cases} 0 & \text{si } t < 0 \\ \frac{t}{\theta} & \text{si } 0 \leq t \leq \theta \\ 1 & \text{si } t > \theta. \end{cases}$$

$\xrightarrow{\bar{\rightarrow}} \left(\frac{t}{\theta}\right)^n$

sea $\varepsilon > 0$, $X_i \sim U(0, \theta)$

$$P(|X_{(n)} - \theta| > \varepsilon) = P(\theta - X_{(n)} > \varepsilon) \\ = P(X_{(n)} < \theta - \varepsilon) = \left(\frac{\theta - \varepsilon}{\theta}\right)^n \xrightarrow{n \rightarrow \infty} 0.$$

Luego $X_{(n)} \xrightarrow{P} \theta$. Luego $X_{(n)}$ es también un estimador razonable para θ . Lo vamos a llamar B_n

¿Cuál de los dos estimadores es mejor?

O sea, ¿cuál de los dos jugadores tiene más chance de ganar?

Vimos que ambos estimadores convergen en probabilidad a θ cuando $x_i \sim U(0, \theta)$. Pero esto solamente da información del comportamiento asintótico. ¿Cuál criterio es relevante cuando n es pequeño (θ fijo), digamos $n=10$?

Propiedades de los estimadores.



(o sea, cualidades, criterios para compararlos)

Sean $x_1, \dots, x_n \sim F_\theta$

Def: Si $E_\theta(\hat{\theta}(x_1, \dots, x_n)) = \theta \quad \forall \theta \in \Theta$, decimos que $\hat{\theta}$ es un estimador insesgado de θ .

Llamamos sesgo de un estimador a la cantidad

$$b(\hat{\theta}, \theta) = E_\theta(\hat{\theta}(x_1, \dots, x_n)) - \theta.$$

Decimos que $\hat{\theta}_n$ es una sucesión de estimadores asintóticamente insesgada (θ , más coloquialmente, que $\hat{\theta}_n$ es un estimador asintóticamente insesgado)

Si

$$E_\theta(\hat{\theta}_n(x_1, \dots, x_n)) \xrightarrow{n \rightarrow \infty} \theta.$$

La insesgadez de un estimador asegura que sus valores están típicamente centrados alrededor del θ con el que fueron producidos. Pero no excluye la posibilidad de que fluctúen mucho alrededor de dicho valor, lo que disminuye su calidad como estimador.

La varianza del estimador mide esta segunda¹² calidad, y por lo tanto es deseable que sea pequeña.

Comparemos A_n y B_n : Si $X_i \sim U(0, \theta)$,

$$E_{\theta}(A_n) = E_{\theta}(2\bar{X}_n) = 2E_{\theta}(X_1) = 2 \cdot \frac{\theta}{2} = \theta.$$

$$E_{\theta}(B_n) =$$

↑

$$F_{B_n}(t) = \frac{t^n}{\theta^n} I_{[0, \theta]}(t) + 1 I_{(\theta, +\infty)}(t)$$

$$\Rightarrow f_{B_n}(t) = \frac{n t^{n-1}}{\theta^n} I_{[0, \theta]}(t)$$

Luego

$$E_{\theta}(B_n) = \int_{-\infty}^{+\infty} t \cdot f_{B_n}(t) dt = \int_0^{\theta} \frac{n}{\theta^n} t^n dt = \left. \frac{n}{\theta^n} t^{n+1} \right|_0^{\theta}$$

$$= \frac{n}{\theta^n} \frac{\theta^{n+1}}{(n+1)} = \frac{n \theta}{n+1}$$

\$A_n\$ es insesgado. El sesgo de \$B_n\$ es \$\left(\frac{n}{n+1}\theta - \theta\right) = \frac{-\theta}{n+1}\$

$$V_{\theta}(A_n) = V_{\theta}(2\bar{X}_n) = 4V(\bar{X}_n) = \frac{4}{n}V(X_1) = \frac{\theta^2}{12n}$$

$V_{\theta}(A_n) = \frac{\theta^2}{3n}$

Para B_n :

$$\begin{aligned} E_{\theta}(B_n^2) &= \int_0^{\theta} t^2 \frac{n t^{n-1}}{\theta^n} dt = \frac{n}{\theta^n} \int_0^{\theta} t^{n+1} dt = \frac{\theta^{n+2}}{\theta^n (n+2)} \Big|_0^{\theta} \\ &= \frac{\theta^{n+2}}{\theta^n} \frac{n}{(n+2)} = \theta^2 \frac{n}{n+2} \end{aligned}$$

$$\boxed{Var_{\theta}(B_n) = \theta^2 \frac{n}{n+2} - \frac{n^2 \theta^2}{(n+1)^2} = \frac{n}{(n+1)^2(n+2)} \theta^2.}$$

Ambas varianzas tienden a 0 con n .

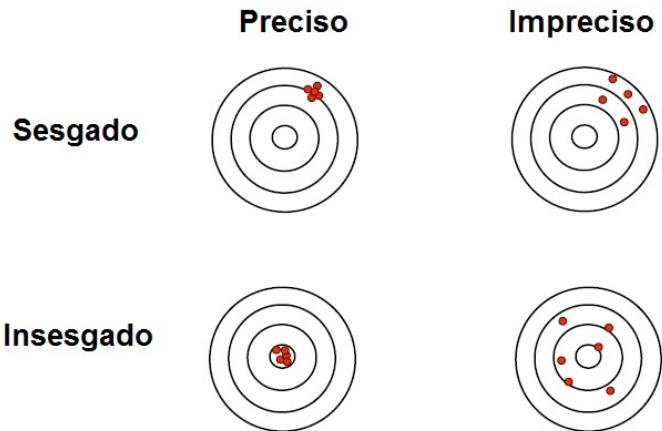
$$\underset{\theta}{V(A_n)} \sim (1/n).$$

$$\underset{\theta}{V(B_n)} \sim (1/n^2) \leftarrow \text{VARÍA MENOS}$$

pero alrededor del número equivocado

Podemos exemplificar Sesgo y precisión con la analogía al tiro al blanco. Pensamos que el blanco es el θ a estimar (descon)

Sesgo vs. Precisión



Tenemos 4 tiradores (o 4 estimadores) con las 4 combinaciones posibles de alto/bajo sesgo y alta/baja dispersión/precisión

En otras palabras, un estimador insesgado evita errores sistemáticos.

¿Cómo balanceamos los 2 requisitos considerados?

- ↳ poco sesgo (θ o nada de sesgo) ✓
- ↳ poca varianza. ✓

Una medida de la calidad de un estimador está dada por el **Error Cuadrático Medio**: (ECM).

$$\text{ECM}(\hat{\theta}, \theta) = E((\hat{\theta} - \theta)^2)$$

En inglés,
Mean Squared
Error (MSE)

Prop: $\text{ECM}(\hat{\theta}, \theta) = \text{Var}(\hat{\theta}) + [b(\hat{\theta}, \theta)]^2$

↑ sesgo (bias)
al cuadrado

dem: Ejercicio. 1

Luego, el criterio para elegir al mejor estimador que suele emplearse es elegir al que tenga menor Error cuadrático medio entre los disponibles.

Por lo que sesgo y varianza deben minimizarse simultáneamente. A veces, es preferible tener un poco de sesgo con tal de achicar el ECM.

Compararemos las propuestas de A y B en términos del ECM:

$$\text{ECM}(A_n, \theta) = V_\theta(A_n) + b^2(A_n, \theta) = \frac{\theta^2}{3n}$$

$$\text{ECM}(B_n, \theta) = V_\theta(B_n) + b^2(B_n, \theta) =$$

$$= \theta^2 \frac{n}{(n+1)^2(n+2)} + \frac{(-1)^2}{(n+1)^2} \theta^2 = \frac{\theta^2}{(n+1)^2} \left[\frac{n}{n+2} + 1 \right]$$

$$= \frac{\theta^2}{(n+1)^2} \frac{2n+2}{n+2} = \frac{2\theta^2}{(n+1)(n+2)}.$$

Para $n=10$, tenemos.

$$\text{ECM}(A_{10}, \theta) = \frac{\theta^2}{30}$$

$$\text{ECM}(B_{10}, \theta) = \frac{\theta^2}{66}$$

Este es preferible

Ejercicio 2: conseguir a B_n para que resulte insesgado, multiplicándolo por la constante adecuada. Llamemos C_n a dicho estimador de θ . Hallar el límite en probabilidad de C_n y calcular su ECM para n arbitrario y para $n=10$. Para $n=10$, ¿cuál estimador de θ es preferible: A_{10} , B_{10} ó C_{10} ? (en términos del ECM).

Otra propiedad importante para los estimadores.

Def: Se dice que la sucesión $T_n = T_n(x_1, \dots, x_n)$ de estimadores de $h(\theta)$ es débilmente consistente si

$$X_i \sim F(\cdot, \theta) \text{ resulta } P_{\theta}(|T_n - h(\theta)| > \epsilon) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \epsilon > 0.$$

Es decir si $T_n \xrightarrow{\text{P}_{\theta}} h(\theta) \quad \forall \theta \in \Theta$

(Se dicen fuertemente consistentes si la convergencia es casi segura.)

En los ejemplos que vimos, los estimadores son débilmente consistentes

¿Cómo encontramos/proponemos un estimador?

↳ A veces un estimador surge de forma intuitiva.

↳ Pero hay métodos para encontrarlos de manera sistemática.

- ↳ Método de máxima verosimilitud ✓
- ↳ Método de momentos.
- ↳ Minimización de una función de pérdida preestablecida (cuad. mínimos, x g).

10 Método de Máxima verosimilitud ¹

10.1 Ejemplo

Suponga que tiene dos monedas en su bolsillo. Una de ellas es equilibrada, mientras que la otra no lo es. La probabilidad de observar cara con la moneda equilibrada es 0.5 mientras que con la no equilibrada la probabilidad de observar cara es 0.8. Identificamos cara y ceca con 1 y 0, respectivamente. Suponga ahora que saca una de las monedas de su bolsillo y deberá determinar si se trata de la moneda equilibrada o de la otra, pudiendo tirarla muchas veces en forma independiente. Suponga que en $n = 100$ lanzamientos observa la muestra

$$\underbrace{1, \dots, 1}_{12 \text{ veces}}, \underbrace{0, \dots, 0}_{5 \text{ veces}}, \underbrace{1, \dots, 1}_{23 \text{ veces}}, \underbrace{0, \dots, 0}_{8 \text{ veces}}, \underbrace{1, \dots, 1}_{15 \text{ veces}}, \underbrace{0, \dots, 0}_{3 \text{ veces}}, \underbrace{1, \dots, 1}_{11 \text{ veces}}, \underbrace{0, \dots, 0}_{4 \text{ veces}}, \underbrace{1, \dots, 1}_{13 \text{ veces}}, \underbrace{0, \dots, 0}_{6 \text{ veces}} \quad (10.1)$$

¿Cuál de las dos monedas diría que está utilizando? Para responder a esta pregunta, vamos a calcular $L(0.8)$ y $L(1/2)$, la probabilidad de observar la muestra (10.1) con cada una de las posibles monedas. Para calcular dichas probabilidades debemos multiplicar la probabilidad de cada una de las observaciones, siendo que la muestra fue generada en forma independiente. Suponiendo que está utilizando la moneda *no* equilibrada, tenemos que

$$L(0.8) = \underbrace{0.8 \dots 0.8}_{12 \text{ veces}} \underbrace{0.2 \dots 0.2}_{5 \text{ veces}} \underbrace{0.8 \dots 0.8}_{23 \text{ veces}} \underbrace{0.2 \dots 0.2}_{8 \text{ veces}} \underbrace{0.8 \dots 0.8}_{15 \text{ veces}} \underbrace{0.22 \dots 0.2}_{3 \text{ veces}} \\ \underbrace{0.8 \dots 0.8}_{11 \text{ veces}} \underbrace{0.2 \dots 0.2}_{4 \text{ veces}} \underbrace{0.8 \dots 0.8}_{13 \text{ veces}} \underbrace{0.2 \dots 0.2}_{6 \text{ veces}} = (0.8)^{74} (0.2)^{26} = 4.5231 \cdot 10^{-26}$$

siendo 74 el número de caras observadas en las $n = 100$ repeticiones. Análogamente, tenemos que

$$L(1/2) = (1/2)^{74} (1/2)^{26} = 7.8886 \cdot 10^{-31},$$

Y ahora, ¿cuál moneda diría usted que está utilizando? La propuesta de máxima verosimilitud consiste en pensar que la moneda que estamos utilizando es aquella que le asigna una probabilidad más alta a la muestra observada (*i.e.* la que hace a la muestra observada más creíble, *más verosímil*). Siendo que

$$L(0.8) > L(1/2)$$

concluimos que está utilizando la moneda no equilibrada.

Supongamos ahora que desconoce por completo qué tipo de monedas tiene en su bolsillo. Siendo que en la muestra de $n = 100$ lanzamientos observó 74 caras, ¿cuál

¹Basado en unas notas sobre estimación escritas por Mariela Sued y M. Eugenia Szretter

diría usted que es la probabilidad de obtener cara en un tiro con la moneda que está utilizando?

Si θ denota la probabilidad de obtener cara en un tiro, siguiendo con el razonamiento empleado en el caso anterior, vamos a elegir aquel valor para el cual la muestra observada tenga mayor probabilidad. Si denotamos por $L(\theta)$ a la probabilidad de observar (10.1), que contiene 74 caras en $n = 100$ repeticiones, tenemos que

$$L(\theta) = \theta^{74}(1 - \theta)^{26}.$$

$L(\theta)$ se llama *función de verosimilitud*. Mide cuál es la probabilidad de observar nuestra realización (nuestros datos) cuando la probabilidad de cara es θ . El estimador de máxima verosimilitud para este modelo está dado por $\hat{\theta}$ que maximiza $L(\theta)$. Siendo el logaritmo una función creciente, la función $L(\theta)$ o su logaritmo, dado por $l(\theta) = \ln(L(\theta))$ se maximizan en el mismo valor. Este último problema suele resultar, analíticamente, más sencillo a la hora de maximizar. En el ejemplo que estamos considerando, tenemos que $l(\theta) = 74 \ln(\theta) + 26 \ln(1 - \theta)$, se maximiza en $\hat{\theta} = 74/100$. Decimos que $74/100$ es el *valor estimado* para θ en base a los datos observados.

Tenemos así que la probabilidad estimada de obtener cara con la moneda que está utilizando es $\hat{\theta} = 0.74$.

10.2 Caso Discreto

Tratemos de generalizar el procedimiento empleado en este ejemplo. Partimos de los datos $\mathbf{x} = (x_1, \dots, x_n)$, una realización de una muestra X_1, \dots, X_n de variables aleatorias discretas donde la distribución asociada a las coordenadas (la misma para todos siendo las variables i.d.) pertenece al modelo $\mathcal{M} = \{p(\cdot, \theta), \theta \in \Theta\}$. En el ejemplo de la moneda, la muestra observada está dada en (10.1), con $n = 100$. La distribución de las variables es Bernoulli, con lo cual el modelo puede ser representado por

$$\mathcal{M} = \{\mathcal{B}(1, \theta), \theta \in [0, 1]\}.$$

Tenemos entonces que el espacio al cual pertenece el parámetro que indexa el modelo está dado por $\Theta = [0, 1]$.

Tal como hicimos en el ejemplo, dada la realización x_1, \dots, x_n de la muestra, la función de verosimilitud da la probabilidad de que la muestra tome el valor observado cuando la distribución que la genera tiene f.p.p. $p(\cdot, \theta)$. Dicho en otras palabras, es el valor de la función de probabilidad puntual del vector (X_1, \dots, X_n) evaluada en $\mathbf{x} = (x_1, \dots, x_n)$ cuando X_1, \dots, X_n son i.i.d. con función de probabilidad puntual $(p \cdot, \theta)$.

Definición 10.1. Función de verosimilitud: Caso discreto. Sea X_1, \dots, X_n una muestra proveniente del modelo \mathcal{M} . Dada la realización $\mathbf{x} = (x_1, \dots, x_n)$ de la muestra, la función de verosimilitud está dada por

$$L(\cdot ; \mathbf{x}) : \Theta \rightarrow [0, 1]$$

$$L(\theta ; \mathbf{x}) = P(X_1 = x_1, \dots, X_n = x_n) = p_{X_1, \dots, X_n}(x_1, \dots, x_n),$$

cuando las variables fueron generadas según la f.p.p. $p(\cdot, \theta)$. Utilizando el hecho de que X_1, \dots, X_n son i.i.d. con función de probabilidad puntual $p(\cdot, \theta)$, concluimos que

$$L(\theta ; \mathbf{x}) = P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n p(x_i, \theta).$$

La propuesta de máxima verosimilitud consiste en asignarle a la la realización $\mathbf{x} = (x_1, \dots, x_n)$ el valor $h_n(x_1, \dots, x_n)$ que maximiza la función de verosimilitud $L(\cdot ; \mathbf{x})$. Es decir, consideramos el valor

$$h_n(x_1, \dots, x_n) \in \Theta \text{ que maximiza } L(\theta, \mathbf{x}).$$

Observación 10.1. El desarrollo hecho para llegar a la propuesta de verosimilitud se basa en que disponemos de datos x_1, \dots, x_n . Una vez que tenemos una fórmula para $h_n(x_1, \dots, x_n)$, reemplazando los valores observados por la muestra X_1, \dots, X_n obtenemos una variable aleatoria. A esta variable aleatoria la denominamos estimador de máxima verosimilitud.

Definición 10.2. Llamamos estimador de máxima verosimilitud para θ a la variable aleatoria

$$\widehat{\theta}_n = h_n(X_1, \dots, X_n).$$

Ejercicio 10.1. Sea X_1, \dots, X_n una muestra aleatoria de una distribución

- a) $Bi(1, \theta)$, $0 \leq \theta \leq 1$.
- b) $\mathcal{P}(\theta)$, $\theta > 0$.
- c) $\mathcal{G}(\theta)$, $0 \leq \theta \leq 1$.

En cada uno de estos casos, encontrar el estimador de máxima verosimilitud de θ .

Ejercicio 10.2. Sea X_1, \dots, X_n una muestra de una distribución discreta que toma los valores y_1, \dots, y_k con probabilidades p_1, \dots, p_k , respectivamente. Obtener el estimador de máxima verosimilitud para los valores p_1, \dots, p_k , especificando el espacio Θ donde viven los parámetros.

10.3 Caso Continuo

Siguiendo el enfoque propuesto en el caso discreto, cuando trabajamos con variables continuas con modelos indexados mediante una parametrización de las posibles funciones de densidad, la función de verosimilitud evaluada en θ está dada por la función de densidad conjunta de la muestra evaluada en los datos observados, θ es el parámetro asociado a la distribución que genera los datos. Mas específicamente, consideremos el modelo

$$\mathcal{M} = \{f(\cdot, \theta) , \theta \in \Theta\}$$

con $f(\cdot, \theta)$ función de densidad.

Dadas los datos (x_1, \dots, x_n) correspondientes a una realización de la muestra X_1, \dots, X_n , la función de verosimilitud asociada a la muestra evaluada en θ esta dada por la densidad conjunta del vector (X_1, \dots, X_n) evaluada en (x_1, \dots, x_n) asumiendo que $X_i \sim f(\cdot, \theta)$.

Definición 10.3. Función de verosimilitud: Caso continuo. Sea X_1, \dots, X_n una muestra proveniente del modelo \mathcal{M} . Dada la realización $\mathbf{x} = (x_1, \dots, x_n)$ de la muestra, la función de verosimilitud está dada por

$$L(\cdot ; \mathbf{x}) : \Theta \rightarrow \mathbb{R}_{\geq 0}$$

$$L(\theta ; \mathbf{x}) = f_{X_1 \dots X_n}(x_1, \dots, x_n),$$

cuando las variables fueron generadas según la densidad $f(\cdot, \theta)$. Utilizando el hecho de que X_1, \dots, X_n son i.i.d. con función de densidad $f(\cdot, \theta)$, concluimos que

$$L(\theta ; \mathbf{x}) = f_{X_1 \dots X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) = \prod_{i=1}^n f(x_i, \theta).$$

La propuesta de máxima verosimilitud consiste en asignarle a la realización (x_1, \dots, x_n) el valor $h_n(x_1, \dots, x_n)$ que maximiza la función de verosimilitud $L(\cdot ; \mathbf{x})$. Es decir, consideramos el valor

$$h_n(x_1, \dots, x_n) \in \Theta \text{ que maximiza } L(\theta, \mathbf{x}).$$

Observación 10.2. El desarrollo hecho para llegar a la propuesta de verosimilitud se basa en que disponemos de una realización x_1, \dots, x_n . Una vez que tenemos una fórmula para $h_n(x_1, \dots, x_n)$, reemplazando los valores observados por la muestra X_1, \dots, X_n obtenemos una variable aleatoria. A esta variable aleatoria la denominamos estimador de máxima verosimilitud.

Definición 10.4. Llamamos estimador de máxima verosimilitud para θ a la variable aleatoria

$$\widehat{\theta}_n = h_n(X_1, \dots, X_n).$$

Ejercicio 10.3. Sea X_1, \dots, X_n una muestra aleatoria de una distribución

- a) $\mathcal{E}(\theta)$, $\theta > 0$.
- b) $\mathcal{U}[0, \theta]$, $0 < \theta$.
- c) $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$.
- d) $\mathcal{N}(\mu, \sigma^2)$

En cada uno de estos casos, encontrar el estimador de máxima verosimilitud de los parámetros que identifican al modelo.

Ejercicio 10.4. Sea X_1, \dots, X_n una muestra aleatoria de una distribución exponencial desplazada, cuya densidad es

$$f(x) = e^{-(x-\theta)} I_{[\theta, \infty)}(x).$$

Encontrar el estimador de máxima verosimilitud de θ .

Finalicemos con una generalización del criterio de máxima verosimilitud. A veces lo que interesa estimar no es el parámetro θ sino una función de él, digamos $t(\theta)$. Entonces, si $\widehat{\theta}_n$ es el estimador de máxima verosimilitud de θ , entonces $t(\widehat{\theta}_n)$ se denomina el estimador de máxima verosimilitud de $t(\theta)$.

Ejercicio 10.5. Para el ejercicio 10.4, hallar un estimador de máxima verosimilitud de

- a) $P(X_1 > 2\theta)$
- b) $E(X_1)$

Intervalos de confianza

Los estimadores puntuales son útiles para tener una primera idea del valor verdadero de una cierta cantidad desconocida, pero tienen el problema de que uno no sabe cuán buenas son las estimaciones. Para tomar en cuenta el azar involucrado, uno no debería calcular un sólo valor, sino una región que dependa del resultado del experimento y que podamos asegurar que contiene al valor verdadero con una alta certeza. Para eso se construyen los **intervalos de confianza** (o **regiones de confianza**).

Definición: Sea $x_1, \dots, x_n \sim F(\cdot, \theta)$ v.a.iid (una muestra aleatoria) con $\theta \in \Theta$. Un intervalo de confianza para el parámetro θ de nivel $1-\alpha$ es un intervalo $C_n = (a(x_1, \dots, x_n), b(x_1, \dots, x_n))$ con extremos que son funciones de la muestra tales que

$$P_\theta (\theta \in C_n) \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

En palabras, $C_n = (a, b)$ contiene a θ con probabilidad $(1-\alpha)$. Llamamos $1-\alpha$ al nivel de confianza (o cobertura) del intervalo de confianza que abreviamos IC.

Conviene Recordar:

- C_n es aleatorio
- θ es fijo

Usualmente se usan intervalos de confianza de nivel 0,95, lo cual corresponde a tomar $\alpha=0,05$.
Algunos textos lo escriben en % (intervalo de confianza de nivel 95%).

Si θ fuera un vector (en \mathbb{R}^k en vez de en \mathbb{R}) usamos una **región de confianza** (como un rectángulo o una elipse) en vez de un intervalo.

CUIDADO: Hay mucha confusión sobre como interpretar un intervalo de confianza. Un intervalo de confianza **NO** es una afirmación probabilística sobre θ pues θ es una cantidad fija, no una r.a.
Entonces, ¿cómo lo interpretamos?

En el día 1 recolectamos datos y construimos un intervalo de confianza de nivel 0,95 para un parámetro θ_1 .

En el día 2, recolectamos nuevos datos y construimos un intervalo de confianza de nivel 0,95 para un nuevo parámetro θ_2 , con datos independientes de los del día 1, para θ_2 que no tiene por qué guardar relación con θ_1 .

En el día 3...

Así seguimos, construyendo una colección de IC para una colección de parámetros $\theta_1, \theta_2, \dots$

Entonces, un 95% de los intervalos construidos atraparán al verdadero parámetro.

Ejemplo: Por experiencia previa se sabe que un método de medición del porcentaje de hierro en un mineral arroja resultados con distribución normal con desvío estándar 0,12. Hacemos 4 determinaciones y obtenemos (en %) los valores:

15,17 15,32 15,46 15,25

Queremos construir un IC de nivel 0,95 para $\mu = E(X_i)$.

¿Cómo lo modelamos? Definimos

X_i = iésima determinación del contenido de hierro en el mineral.

$X_i \sim N(\mu, \sigma^2)$ con $\sigma = 0,12$ vaid

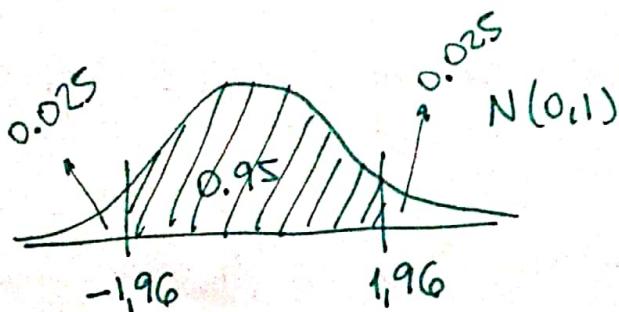
El estimador de μ : $\hat{\mu} = \bar{x}_n$

Sabemos que $\bar{x}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

$$\Leftrightarrow (\bar{x}_n - \mu) \frac{1}{\sigma} \sim N(0,1)$$

función de la muestra y del parámetro, en la que no aparecen términos desconocidos

↓ conocemos que la distribución, no depende del parámetro desconocido.



De tabla:

$$0.95 = P\left(-1.96 \leq \frac{(\bar{x}_n - \mu)}{\sigma} \leq 1.96\right)$$

$$= P\left(-1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{x}_n - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(-\bar{x}_n - 1.96 \frac{\sigma}{\sqrt{n}} \leq -\mu \leq 1.96 \frac{\sigma}{\sqrt{n}} - \bar{x}_n\right)$$

$$= P\left(\bar{x}_n + 1.96 \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x}_n - 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

Luego un intervalo de confianza de nivel 0.95 para μ es

μ = Contenido medio esperado de hierro en el mineral.
parametro de interés.

$$\left[\bar{x}_n - 1,96 \frac{\sigma}{\sqrt{n}} ; \bar{x}_n + 1,96 \frac{\sigma}{\sqrt{n}} \right].$$

Para la muestra obtenida, tenemos:

$$\bar{x}_n = 15,3 \quad n = 4$$

$$\sigma = 0,1$$

$$\left[15,3 - 1,96 \cdot \frac{0,12}{\sqrt{4}}, 15,3 + 1,96 \cdot \frac{0,12}{\sqrt{4}} \right]$$

$\underbrace{\phantom{1,96 \cdot 0,12 / \sqrt{4}}}_{0,1176}$

Redondeando 0,12

$$[15,18 ; 15,42]$$

¿Qué significa?

¿Significa que $P(\mu \in [15,18 ; 15,42]) = 0,95$?

¡No! No hay nada aleatorio en el evento A!

Por lo que ese evento tiene probabilidad 0 ó 1.

¿Y qué es 0,95, entonces?

Si repitiéramos

todo el experimento

muchas veces.

extraer 1 muestra x_1, \dots, x_n de tamaño n y armar el IC basado en ella; $\bar{x}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

entonces el 95% de los intervalos contendrían al verdadero valor del parámetro de la población (en este caso, la media μ)

Muestra 1: $X_1^{(1)}, \dots, X_n^{(1)}$ → Calculo $\bar{X}_n^{(1)}$ e IC ————— ()

Muestra 2: $X_1^{(2)}, \dots, X_n^{(2)}$ → Calculo $\bar{X}_n^{(2)}$ e IC ————— ()

Muestra 3: $X_1^{(3)}, \dots, X_n^{(3)}$ → Calculo $\bar{X}_n^{(3)}$ e IC ————— ()

⋮
⋮

Muestra B: $X_1^{(B)}, \dots, X_n^{(B)}$ → Calculo $\bar{X}_n^{(B)}$ e IC ————— ()

μ
↑
 μ Verdadero

¿Cuántos de estos
intervalos contendrán
al verdadero μ ?

↳ ¿qué proporción ($\sigma\%$) de ellos lo contendrá?
↳ (Aprox) El 95% de ellos

Pero, en la práctica, nosotros solamente extraemos
1 muestra de tamaño n , que da un intervalo.

$$[15, 18 ; 15, 42]$$

¿En cuál de las situaciones estaremos?

¿Contendrá o no al verdadero μ ? No lo sabemos.

La ciencia trabaja como si lo contuviera.

¿Qué pasaría si tomamos nivel 0,99 en vez de 0,95?

Ejercicio.

- b) Hallar el tamaño de muestra ^{necesario} para que la longitud del IC₁ sea ≤ 0,10.
de nivel 0,95.

$$\text{Busco } n \mid L = 2 \cdot z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq 0,10 \quad (\Rightarrow)$$

$$\Leftrightarrow \frac{2 \cdot z_{\frac{\alpha}{2}} \sigma}{0,10} \leq \sqrt{n} \quad (\Rightarrow) \quad n \geq \frac{4 z_{\frac{\alpha}{2}}^2 \sigma^2}{0,10^2} = \frac{4 \cdot 1,96 \cdot 0,12^2}{0,10^2}$$

$$n \geq 22,13 \rightsquigarrow \text{Rta } n=23 (\text{o más determinaciones})$$

Volvamos. ¿Por qué podemos hallar el IC?

Usamos el resultado

que nos da una expresión pivote:

$$Z = \frac{(\bar{X}_n - \mu)}{\sigma} \sqrt{n} \sim N(0,1)$$

1º) Va que depende de la muestra
y del único parámetro ^{en} cuyo IC
estamos _{desconocido} interesados.

2º) Conocemos la distribución,
que no depende de ningún parámetro _{desconocido}.

Intervalos de Confianza Asintóticos

Def: Dada $X_1, \dots, X_n \sim F(\cdot, \theta)$ una muestra aleatoria, $\theta \in \mathbb{R}$. Diremos que $C_n = (a(X_1, \dots, X_n), b(X_1, \dots, X_n))$ es un intervalo asintótico de nivel $1-\alpha$ para $\theta \in \mathbb{R}$ cuando:

$$\liminf_{n \rightarrow \infty} P_\theta(\theta \in C_n) \geq 1 - \alpha \quad \forall \theta \in \mathbb{R}.$$

Entonces, el Teorema Central del Límite junto con el Teorema de Slutsky nos proporcionan expresiones pivotales para calcular IC de nivel asintótico.

Teo Central del Límite: $(X_i)_{i \geq 1}$ vaid con $E(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2 < \infty$, luego:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0, 1).$$

Muchas veces, σ^2 , la varianza, es desconocida.

Podemos estimarla usando la varianza muestral (en base a las mismas observaciones).

Recordemos:

Varianza muestral

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Vemos que es un estimador consistente de σ^2 , es decir:

$$S_n^2 \xrightarrow{P} \sigma^2$$

Equivalentemente,

$$\frac{S_n^2}{\sigma^2} \xrightarrow{P} 1 \Leftrightarrow \frac{S_n}{\sigma} \xrightarrow{P} 1 \Leftrightarrow \frac{\sigma}{S_n} \xrightarrow{P} 1.$$

Usemos el Teorema de Slutsky:

Sean $(W_n)_{n \geq 1}$, $(V_n)_{n \geq 1}$ sucesiones de va tales que:

$$W_n \xrightarrow{D} W \text{ y } V_n \xrightarrow{P} c \in \mathbb{R}.$$

Entonces:

$$a) W_n + V_n \xrightarrow{D} W + c$$

$$b) V_n \cdot W_n \xrightarrow{D} cW.$$

Además, sabemos que si g es continua, entonces:

$$g(W_n) \xrightarrow{D} g(W)$$

$$\left\{ \begin{array}{l} \frac{(\bar{X}_n - \mu)\sqrt{n}}{\sigma} \xrightarrow{\mathcal{D}} N(0,1) \\ \frac{\sigma}{S_n} \xrightarrow{P} 1 \end{array} \right. \quad \Rightarrow \text{ Slutsky ii)}$$

$$\left\{ \begin{array}{l} \frac{(\bar{X}_n - \mu)\sqrt{n}}{\sigma} \xrightarrow{\mathcal{D}} N(0,1) \\ \frac{\sigma}{S_n} \xrightarrow{n \rightarrow \infty} 1 \end{array} \right.$$

$$\left\{ \begin{array}{l} \frac{(\bar{X}_n - \mu)\sqrt{n}}{S_n} \xrightarrow[n \rightarrow \infty]{\mathcal{D}} N(0,1) \end{array} \right. \quad \boxed{\text{(A)}}$$

asintótico.

Expresión pivote para un IC de nivel $1-\alpha$ para μ .

IC de nivel asintótico $1-\alpha$ para $\mu = E(x_i)$

$X_1, \dots, X_n \rightsquigarrow \text{vaiid}$ con $E(X_i) = \mu$ y $\text{Var}(X_i) = \sigma^2$

Usemos **(A)**

$$1-\alpha = \lim_{n \rightarrow \infty} P\left(-Z_{\alpha/2} < \frac{(\bar{X}_n - \mu)\sqrt{n}}{S_n} < Z_{\alpha/2}\right)$$

$$= \lim_{n \rightarrow \infty} P\left(\bar{X}_n - \frac{S_n}{\sqrt{n}} Z_{\alpha/2} \leq \mu \leq \bar{X}_n + \frac{S_n}{\sqrt{n}} Z_{\alpha/2}\right).$$

Otra aplicación

IC para $\theta = P(\text{éxito})$ en el caso Binomial

asintótico

Sean $X_1, \dots, X_n \sim Bi(1, \theta)$.

Sabemos que:

$$\frac{(\bar{X}_n - \theta)\sqrt{n}}{\sqrt{\theta(1-\theta)}} \xrightarrow[n \rightarrow \infty]{D} N(0,1)$$

por el TCL.

$$E(\bar{X}_n) = E(X_1) = \theta.$$

$$\text{Var}(\bar{X}_n) = \frac{\theta(1-\theta)}{n}$$

Estimador de θ : $\hat{\theta}_n = \bar{X}_n \xrightarrow[n \rightarrow \infty]{\text{LGN}} \theta$.

$$g(\theta) = \sqrt{\theta(1-\theta)} \text{ es continua.}$$

Luego, porque la convergencia en prob es preservada por funciones continuas, tenemos

$$g(\hat{\theta}_n) \xrightarrow{P} g(\theta).$$

$$\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)} \xrightarrow{P} \sqrt{\theta(1-\theta)}$$

o equivalentemente

$$\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{\theta(1-\theta)}} \xrightarrow{P} 1 \Rightarrow \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow{P} 1.$$

pues las funciones continuas preservan la convergencia en probabilidad, y $g(t) = \frac{1}{t}$ es cont

Luego:

$$\frac{(\bar{x}_n - \theta)\sqrt{n}}{\sqrt{\theta(1-\theta)}} \xrightarrow{\mathcal{D}} N(0,1) \quad \text{TCL.} \quad \Rightarrow$$

$$\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow{P} 1.$$

$$\Rightarrow \text{Slutzky b)} \quad \frac{(\bar{x}_n - \theta)\sqrt{n}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} = \boxed{\frac{(\bar{x}_n - \theta)\sqrt{n}}{\sqrt{\bar{x}_n(1-\bar{x}_n)}}} \xrightarrow{\mathcal{D}} N(0,1)$$

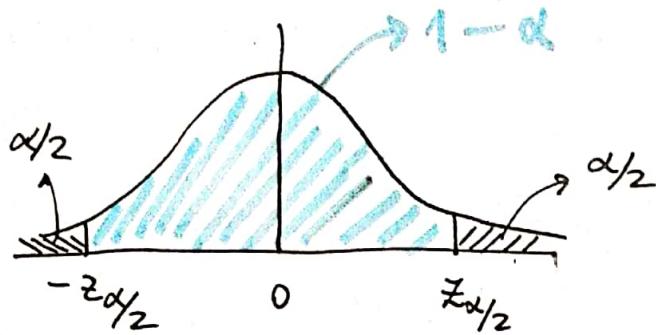
expresión pivote para el IC de nivel asintótico $1-\alpha$ para θ , el parám de una Bernoulli

$$\lim_{n \rightarrow \infty} P\left(-z_{\alpha/2} \leq \frac{(\bar{x}_n - \theta)\sqrt{n}}{\sqrt{\bar{x}_n(1-\bar{x}_n)}} \leq z_{\alpha/2}\right) =$$

$$= \lim_{n \rightarrow \infty} P\left(-\bar{x}_n - z_{\alpha/2} \frac{\sqrt{\bar{x}_n(1-\bar{x}_n)}}{\sqrt{n}} \leq -\theta \leq z_{\alpha/2} \frac{\sqrt{\bar{x}_n(1-\bar{x}_n)}}{\sqrt{n}} - \bar{x}_n\right)$$

$$= \lim_{n \rightarrow \infty} P\left(\bar{x}_n - z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \leq \theta \leq \bar{x}_n + z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}\right)$$

$= 1 - \alpha$ si $Z_{\alpha/2}$ es tal que



Podemos utilizar el IC para Bernoullis en el ejemplo de la encuesta de despenalización del aborto (realizada por Amnistía Internacional en marzo 2012)

Sea

$$X_i = \begin{cases} 1 & \rightarrow \text{si el } i\text{\'esimo encuestado est\'a a favor} \\ & \text{de la ley de despenalizaci\'on} \\ 0 & \rightarrow \text{sino.} \end{cases}$$

$$1 \leq i \leq n = 1600. \quad X_i \sim Be(\theta) \text{ con } X_i \text{ iid.}$$

Sea $\theta = P(X_i = 1)$

parametro

VERDADERA PROPORCI\'ON p\'ublica a favor de la despenalizaci\'on.

$$\hat{\theta} = \bar{x}_n.$$

Su valor observado en la muestra fue

estimador

$$\hat{\theta}_{\text{observado}} = \frac{930}{1600} = 0,58125.$$

estimaci\'on

Un IC de nivel asint\'otico 0,95 para θ basado en la muestra es:

percentil 0,975 de una $N(0,1)$.

$$\left[0,5812 \pm \frac{1,96}{\sqrt{1600}} \sqrt{\frac{930}{1600} \times \frac{1600-930}{1600}} \right] = [0,557 ; 0,605]$$

Se usan en muchas aplicaciones.

Por ejemplo, si queremos comparar 2 tratamientos médicos,

↳ muestra 1 con pacientes bajo Tratamiento 1.

X_i = respuesta del i ésimo paciente bajo T1.
(presión/colesterol/nivel de azúcar...)

$$1 \leq i \leq n_1 \quad \mu_1 = E(X_i) \quad X_1, \dots, X_{n_1} \text{ iid.}$$

→ muestra 2 con pacientes bajo Tratamiento 2

Y_j = respuesta del j ésimo paciente bajo T2.

$$1 \leq j \leq n_2 \quad \mu_2 = E(Y_j) \quad Y_1, \dots, Y_{n_2} \text{ iid}$$

Queremos un IC para $\mu_1 - \mu_2$ basado en las 2 muestras independientes $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$

↳ Receta: encontrar una expresión pivote a partir del estimador $\bar{X}_{n_1} - \bar{Y}_{n_2}$. (asintótico).

Hay otros métodos de construcción de IC.

↳ Usar estimadores de máxima verosimilitud

↳ Usar bootstrap u otras técnicas de remuestreo o simulación.

...

Procesos de Poisson

Muchos procesos que surgen naturalmente cambian su valor en cualquier instante de tiempo real.

Tal proceso se modela matemáticamente con una familia $\{X_t : t \geq 0\}$ de variables aleatorias que toman valores en un espacio S . La teoría general de estos objetos se denomina procesos de tiempo continuo (o procesos estocásticos a tiempo continuo). y es sumamente útil y profunda. Acá sólo miraremos un poco el ejemplo más difundido, el de los procesos de Poisson.

Para definirlo, pensemos que queremos modelar la cantidad de partículas que emite una fuente radioactiva en el tiempo. O la llegada de pedidos a un servidor, o a una cola de atención de reclamos.

Aquí $S = \mathbb{N}_0$ y pensamos que N_t es el número de partículas emitidas en el intervalo de tiempo $(0, t]$ o bien, la cantidad de pedidos recibidos en ese intervalo de tiempo. N_t será entonces el número de ocurrencias/emisiones/llegadas. N_t es un caso particular de proceso de conteo.

¿Qué condiciones le pediremos a N_t que satisfaga?

Sea $N = \{N_t : t \geq 0\}$ una colección de r.a. que toma valores en \mathbb{N}_0 .

P1: $N_0 = 0$, si $s < t$ entonces $N_s \leq N_t$.

(o sea comenzamos con el contador en 0 y a medida que pasa el tiempo sólo puede aumentar la cantidad de ocurrencias).

P2: Homogeneidad: La distribución del número de ocurrencias en un cierto intervalo de tiempo sólo depende de la longitud del mismo.

$$P(N_{t+s} - N_t = j) = P(N_s = j) \quad \forall s, t \geq 0, \forall k \in \mathbb{N}_0$$

($N_{s+t} - N_t$ = número de ocurrencias en $(t, s+t]$).

P3: Independencia: La cantidad de ocurrencias en intervalos de tiempo disjuntos son variables aleatorias independientes. Es decir si $t_1 < t_2 < \dots < t_j$ tenemos que:

$$P(\{N_{t_1} = k_1\} \cap \{N_{t_2} - N_{t_1} = k_2\} \cap \dots \cap \{N_{t_j} - N_{t_{j-1}} = k_j\})$$

$$= P(N_{t_1} = k_1) P(N_{t_2} - N_{t_1} = k_2) \dots P(N_{t_j} - N_{t_{j-1}} = k_j)$$

para todo $k_1, \dots, k_j \in \mathbb{N}_0$, $j \in \mathbb{N}$.

P4. Existe $\lambda > 0$ tal que:

$$P(N_h = 1) = \lambda h + o(h) \quad y$$

$$P(N_h \geq 2) = o(h)$$

donde $o(h)$ cumple que $\lim_{h \rightarrow 0} \frac{o(h)}{h} = 0$

(Observemos que esto dice que
 $P(N_h = 0) = 1 - \lambda h + o(h)$)

Probamos ahora el resultado que le da el nombre
 al Proceso de Poisson.

Teorema 11.1

Para cualquier proceso $(N_t)_{t \geq 0}$
 satisfaciendo P1 - P4 vale que $N_t \sim \mathcal{P}(\lambda t)$.

Definición: En tal caso, decimos que $(N_t)_{t \geq 0}$
 es un Proceso de Poisson con intensidad o parámetro
 λ y lo notamos $(N_t)_{t \geq 0} \sim \text{P.P.}(\lambda)$.

dem: Condicionemos N_{t+h} al valor de N_t . $t, h > 0$
 $j \in \mathbb{N}$

$$P(N_{t+h} = j) = \sum_{i=0}^j P(N_t = i, N_{t+h} = j)$$

$$= \sum_{i=0}^j P(N_t = i, N_{t+h} - N_t = j-i)$$

$$\stackrel{\substack{\uparrow \\ \text{indep}}}{=} \sum_{i=0}^j P(N_t = i) \cdot P(N_{t+h} - N_t = j-i)$$

$$\stackrel{\substack{\uparrow \\ \text{homogeneidad}}}{=} \sum_{i=0}^j P(N_t = i) P(N_h = j-i)$$

$$\stackrel{\substack{\uparrow \\ \text{P4}}}{=} P(N_t = j) \cdot P(N_h = 0) + P(N_t = j-1) P(N_h = 1) + \sum_{i=2}^j P(N_t = j-i) P(N_h = i).$$

$$= (1-\lambda h) P(N_t = j) + \lambda h P(N_t = j-1) + \theta(h)$$

Luego, si llamamos

$$a_j(t) = P(N_t = j)$$

tenemos

$$a_j(t+h) = a_j(t)(1-\lambda h) + a_{j-1}(t)\lambda h + \theta(h), j \neq 0$$

(A)

j=0

PP5

$$\begin{aligned} P(N_{t+h}=0) &= P(N_t=0, N_{t+h}=0) \\ &= P(N_t=0) P(N_{t+h}-N_t=0) \\ &= P(N_t=0) \underbrace{P(N_h=0)}_{1-\lambda h + \theta(h)} \end{aligned}$$

$$a_0(t+h) = a_0(t)(1-\lambda h) + \theta(h) \quad (3)$$

Tratamos de escribir una ecuación diferencial para a_j :

(copiamos) $(j \geq 1)$ (A)

$$a_j(t+h) = a_j(t)(1-\lambda h) + a_{j-1}(t)\lambda h + \theta(h).$$

Restamos $a_j(t)$ y dividimos por h :

$$\frac{a_j(t+h) - a_j(t)}{h} = \frac{-\lambda h a_j(t) + a_{j-1}(t)\lambda h + \theta(h)}{h}$$

Tomamos $\lim_{h \rightarrow 0}$ y obtenemos:

$$\begin{cases} a'_j(t) = \lambda (a_{j-1}(t) - a_j(t)) & \text{si } j \neq 0 \\ a'_0(t) = \lambda a_0(t) & \end{cases} \quad (1)$$

$$(2)$$

La condición de frontera es: $a_j(0) = \begin{cases} 1 & \text{si } j=0 \\ 0 & \text{sino} \end{cases}$

Esto es una colección de ecuaciones diferenciales -en diferencias, para las $a_j(t)$. Resolvemos por inducción.

Resolvemos (2) sujeto a la condición inicial

$$a_0(0) = 1$$

y obtenemos:

$$a_0(t) = e^{-\lambda t} \quad (t > 0).$$

Reemplazamos en (1) con $j=1$ para obtener:

$$a_1'(t) = \lambda e^{-\lambda t} - a_1(t)\lambda.$$

Luego $a_1'(0) = 0$

$$\downarrow$$

$$a_1(t) = \lambda t e^{-\lambda t}$$

y continuando, se puede probar por inducción

que
$$a_j(t) = \frac{(\lambda t)^j}{j!} e^{-\lambda t} \quad j=0, 1, \dots$$



Ejemplo: Un famoso experimento llevado a cabo por Rutherford, Chadwick y Ellis publicado en 1920 (sacado del libro Breiman, L. "Probability and Stochastic Processes , with a view toward Applications")

Se registran el número de partículas emitidas por un material radioactivo masivo durante 2608 intervalos de tiempo de 7,5 segundos de duración, cada uno.

Sea $R(k)$ la proporción de intervalos de tiempo en los que se observaron k partículas.

k	$R(k)$	$P(k)$
0	0,024	0,021
1	0,078	0,081
2	0,147	0,156
3	0,201	0,201
4	0,202	0,195
5	0,157	0,151
6	0,105	0,097
7	0,053	0,054
8	0,017	0,026
9	0,010	0,011
≥ 10	0,006	0,007

$P(k) = P_W(k)$ con $W \sim P(\lambda = 3,87)$. Vemos que ambas columnas tienen valores muy parecidos.

Ejemplo 8 Ejercicio 1: Al analizar el patrón con el que los alemanes bombardaron la ciudad de Londres durante la segunda guerra Mundial, los investigadores partieron el área, en 576 sectores pequeños, cada uno de $\frac{1}{4}$ de km^2 . Los investigadores partieron el área, en 576 sectores pequeños, cada uno de $\frac{1}{4}$ de km^2 .

La siguiente tabla es parte de lo observado:

k	0	1	2	3	4	5 y más
Nº de sectores	229	211	93	35	7	1

- ¿Bajo qué condiciones sería esperable que una distribución Poisson sea un buen modelo para los datos sobre el número de bombas caídas en cada sector?
- Ajuste una distribución Poisson a los datos estimando a λ por el promedio (tome "5 y más" como 5) de ataques de bomba recibido por sector.
- Use este $\hat{\lambda}$ para comparar probabilidades estimadas con frecuencias observadas (como en el ejemplo de las emisiones radioactivas) e, multiplicando por 576, compare la cantidad de sectores observados para cada k con los esperados. (Debería obtenerse: 228, 212, 98, 30, 7, 1)

Este ejemplo está en Feller, W. *Introduction to Probability Theory and its Applications*, Vol I, y este ejercicio está tomado del libro de Breiman L citado antes.

Hay una **formulación alternativa** y **equivalente** del Proceso de Poisson que puede proporcionar cierta claridad sobre su funcionamiento. Sean T_0, T_1, \dots dados por:

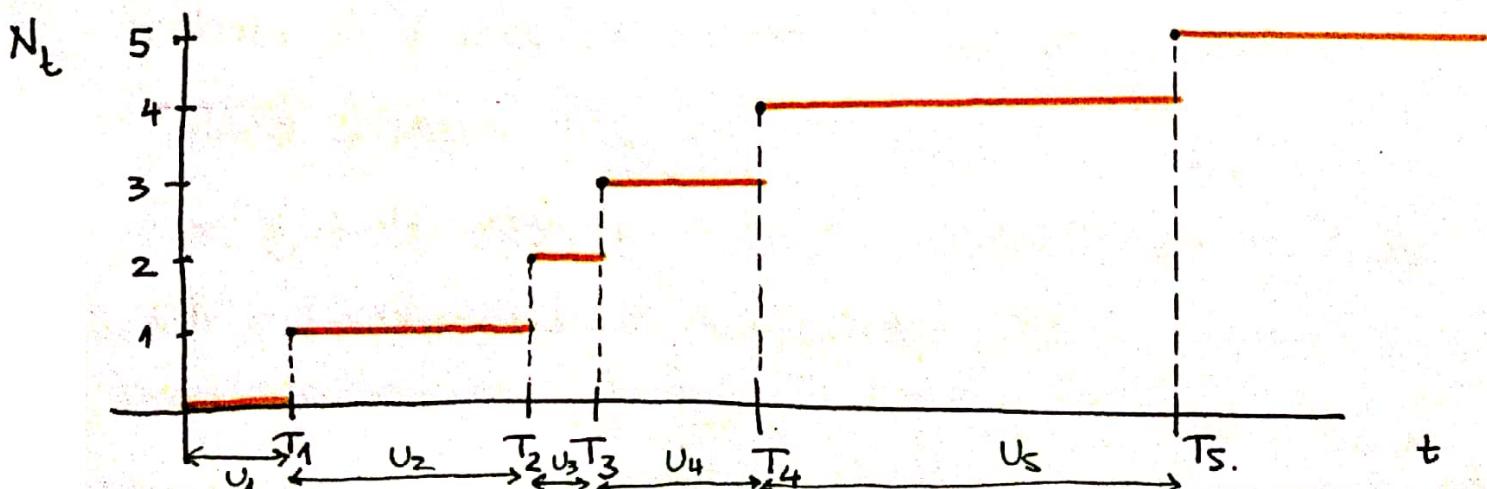
$$T_0 = 0 \quad T_k = \inf \{ t \in \mathbb{R} : N_t = k \} \quad (3)$$

Entonces, T_k es el tiempo en el que se produce el k -ésimo arribo/llegada/ocurrencia. Los tiempos entre eventos U_1, U_2, \dots dados por

$$U_k = T_k - T_{k-1} \quad (4) \quad \underbrace{\text{(el proceso)}}$$

Son variables aleatorias. A partir de conocer a N podemos encontrar U_1, U_2, \dots a partir de (3) y (4). Recíprocamente, podemos reconstruir a N a partir de los $(U_i)_i$ por:

$$T_k = \sum_{i=1}^k U_i, \quad N_t = \max \{ n : T_n \leq t \}$$



Teorema 11.2 Sea $N_t \sim \text{PP}(\lambda)$.

Entonces los tiempos entre eventos U_1, \dots, U_k, \dots

son independientes y cada uno tiene distribución $E(\lambda)$.

dem:

Basta ver que U_1, \dots, U_k son indep y tienen la dist $E(\lambda) \forall k \in \mathbb{N}$.

1º) Veamos que (T_1, \dots, T_k) tiene densidad conjunta

$$f(t_1, \dots, t_k) = \lambda^k e^{-\lambda t_k} \quad 0 < t_1 < \dots < t_k.$$

siendo

$$T_k = \inf \{t : N_t = k\}$$

el tiempo en el que ocurre el k -ésimo arribo.

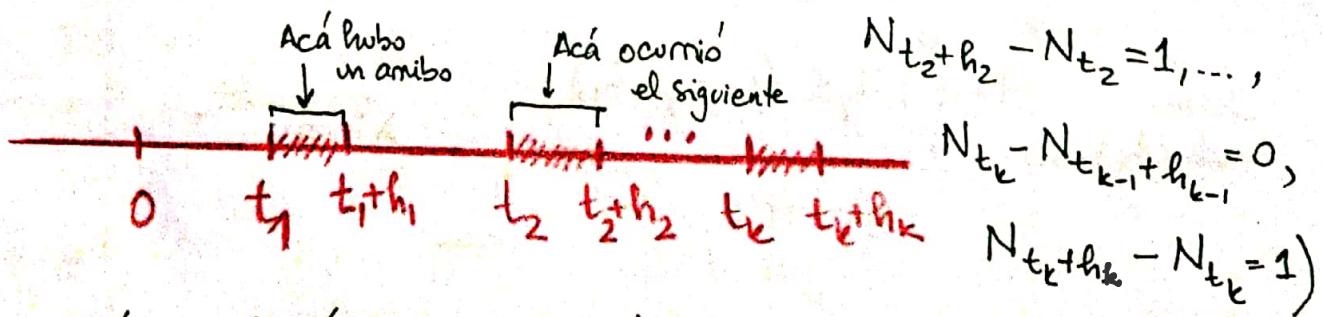
Tenemos

$t_1 < \dots < t_k$. y h_1, \dots, h_k positivos
 $h_i > 0$ pero pequeños de modo que los intervalos
 $\{(t_i, t_i + h_i] \mid i = 1, \dots, k\}$

sean disjuntos. Entonces:

$$P(T_1 \in (t_1, t_1 + h_1], \dots, T_k \in (t_k, t_k + h_k]).$$

$$= P(N_{t_1} = 0, N_{t_1 + h_1} - N_{t_1} = 1, N_{t_2} - N_{t_1 + h_1} = 0,$$



$$= P(N_{t_1} = 0) \cdot P(N_{t_1 + h_1} - N_{t_1} = 1) \cdot P(N_{t_2} - N_{t_1 + h_1} = 0) \cdot P(N_{t_2 + h_2} - N_{t_2} = 1) \cdots$$

$$\text{int. disjuntos.} \therefore \text{r.a. indep} \quad \cdots P(N_{t_k} - N_{t_{k-1} + h_{k-1}} = 0) \cdot P(N_{t_k + h_k} - N_{t_k} = 1)$$

$$= P(N_{t_1} = 0).$$

$$\prod_{j=1}^k P(N_{h_j} = 1) \cdot \prod_{j=2}^k P(N_{t_j - t_{j-1} - h_{j-1}} = 0)$$

= (A)

$$N_{t_1 + h_1} - N_{t_1} \sim N_{h_1} \sim P(\lambda h_1)$$

$$N_{t_j + h_j} - N_{t_j} \sim N_{h_j} \sim P(\lambda h_j)$$

$$N_{t_2} - N_{t_1 + h_1} \sim N_{t_2 - t_1 - h_1} \sim P((t_2 - t_1 - h_1)\lambda)$$

(Es el producto de Probabilidades de v.a. Poisson)

$$W \sim P(\lambda) \Rightarrow p_w(n) = \frac{e^{-\lambda} \lambda^n}{n!} \quad \forall n \geq 0$$

En particular:

$$p_w(0) = e^{-\lambda}$$

$$p_w(1) = e^{-\lambda} \lambda.$$

Luego:

$$(A) = \prod_{j=1}^k e^{-\lambda h_j} j(\lambda h_j) \prod_{j=2}^k e^{-\lambda(t_j - t_{j-1} - h_{j-1})} \cdot e^{-\lambda t_1}.$$

$$= e^{-\lambda \left[\sum_{j=1}^k h_j + \sum_{j=2}^k (t_j - t_{j-1}) - \sum_{j=2}^k h_{j-1} + t_1 \right]} \lambda^k \cdot \prod_{j=1}^k h_j$$

$$= e^{-\lambda h_k - \lambda t_k} \lambda^k \prod_{j=1}^k h_j$$

$$\lim_{\max(h_i) \rightarrow 0} \frac{P\left(\bigcap_{j=1}^k T_j \in (t_j, t_j + h_j]\right)}{\prod_{j=1}^k h_j} = \lambda^k e^{-\lambda t_k}$$

2º) Ahora, hacemos un cambio de variable.

Sea $g: (0, +\infty)^k \rightarrow (0, +\infty)^k$ dada por

$$g(t_1, \dots, t_k) = (t_1, t_2 - t_1, \dots, t_k - t_{k-1}).$$

demodo que $(v_1, \dots, v_k) = g(T_1, \dots, T_k)$.

↑
tiempos entre
ambos

↑
tiempos de los
ambos o llegadas.

Luego g es inversible

$$A = \{(t_1, \dots, t_k) : 0 < t_1 < t_2 < \dots < t_k\}$$

$$g(A) = (0, +\infty)^k$$

La inversa $g^{-1}: (0, +\infty)^k \rightarrow A$.

$$g^{-1}(v_1, \dots, v_k) = (v_1, v_1 + v_2, \dots, v_1 + \dots + v_k).$$

Con Jacobiano $Jg^{-1} = 1$ (matriz diferencial es triangular, con unos en el Δ superior)

Luego

$$\begin{aligned} f_{V_1, \dots, V_k}(v_1, \dots, v_k) &= f_{T_1, \dots, T_k}(g^{-1}(v_1, \dots, v_k)) |Jg^{-1}(v_1, \dots, v_k)| \\ &= \lambda^k \cdot e^{-\lambda \sum_{i=1}^k v_i} \cdot \underbrace{\prod_{i=1}^k 1_A(g^{-1}(v_1, \dots, v_k))}_{\prod_{i=1}^k 1_{(0, +\infty)}(v_i)} = \prod_{i=1}^k \lambda e^{-\lambda v_i} I_{(0, +\infty)}(v_i) \end{aligned}$$



El Teorema 2 permite dar una manera de construir o generar observaciones de un Proceso de Poisson de intensidad λ . Si se quiere generar un PP en $[0, T]$, se pueden generar n independientes i.i.d. con distribución $\text{Exp}(\lambda)$ y a partir de ellas construir T_n y N_t :

$$U_1, U_2, \dots \sim \text{Exp}(\lambda) \text{ iid.}$$

$$T_n = \sum_{i=1}^n U_i \sim P(n, \lambda), \quad T_0 = 0.$$

Finalmente

$$N_t = \max \{ n : T_n \leq t \} = \sum_{k \geq 1} I_{(0, t]} (T_k)$$

Una demostración de que este método funciona puede verse en Durrett, R (2019) "Probability: Theory and Examples", Sección 3.7. Del apunte de Pablo Ferrari, para Proba (M).

No es la única forma de generar un Proceso de Poisson, vale también el siguiente resultado:

En un PP(λ), la $P(T_1, \dots, T_k \in B \mid N_t = k) = P(X_1, \dots, X_k \in B)$ donde $X_1, \dots, X_k \sim U(0, t)$ independientes, B Boreliano de \mathbb{R}^k
 Es decir que si $N_t = k$ la posición de los puntos ^(arribas) es la misma que la de ~~los~~ uniformes en $[0, t]$. Este resultado permite dar otra construcción de los PP que además puede extenderse a \mathbb{R}^k (tomaendo uniformes en rectángulos en vez de intervalos).

Una cualidad interesante de los PP es que permiten introducir la distribución Exponencial, Gamma y Poisson de manera natural e interrelacionada.

La falta de memoria de la distribución exponencial es la clave para que estos procesos estén bien definidos: no importa desde cuándo han estado funcionando los mecanismos que producen emisiones/arribos/ocurrencias según un PP:

Si $X \sim \text{Exp}(\lambda)$, sabemos que $X|_{X>t} \sim \text{Exp}(\lambda)$

por lo tanto, si comenzamos a estudiar un PP que "lleva un tiempo t " de iniciado (comenzamos a registrar la emisión de partículas radiactivas en algún momento), la dist del proceso sera la de un PP.

El hecho de que el mínimo entre 2 o más variables con distribución exponencial también sea una v.a. con distribución exponencial está detrás del hecho de que la suma de dos PP también lo es. Terminamos con dos ejercicios. más

Ejercicio 2: Los eventos (las llegadas) suceden de acuerdo con un PP con tasa $\lambda = 3$ por hora.

- Hallar la probabilidad de que no ocurra ningún evento (no llegue nadie) entre las 8 y las 10 de la mañana.
- ¿Cuál es el valor esperado de eventos que ocurren entre las 8 y las 10 de la mañana?
- ¿Cuál es el horario esperado para la ocurrencia del quinto evento después de las 2 PM?

Ejercicio 3 (Suma o superposición).

Sean $S \sim PP(\lambda)$ y $T \sim PP(\mu)$ independientes.

Probar que $N = S + T$ es un PP. ¿De qué parámetro? (Sug: probar que N satisface las 4 propiedades).

Observar que esto corresponde a contar emisiones de 2 ^{fuentes o} tipos distintos, o ocurrencias/llegadas de 2 (o más) categorías distintas.