

Espacios de Probabilidad
Elementos de Análisis Combinatorio
(Borradores, Curso 23)

Sebastian Grynberg

11-13 de marzo 2013



Andrei Nikolaevich Kolmogorov (1903-1987)
Estableció los fundamentos de la Teoría de Probabilidad en 1933

*“se aprende a pensar abstractamente
mediante el pensamiento abstracto.”*

(G.W.F. Hegel)

Índice

1. Teoría general	3
1.1. Los axiomas de Kolmogorov	3
1.2. Relación con los datos experimentales	5
1.3. Corolarios inmediatos de los axiomas	7
1.4. Sobre el axioma de continuidad	7
1.5. σ -álgebras y teorema de extensión	10
2. Simulación de experimentos aleatorios con espacio muestral finito	11
2.1. Números aleatorios	11
2.2. Simulación de experimentos aleatorios	12
2.3. Estimación de probabilidades	13
3. Elementos de Análisis Combinatorio	17
3.1. Regla del Producto	17
3.2. Muestras ordenadas	18
3.3. Subpoblaciones	21
3.4. Particiones	23
3.5. Distribución Hipergeométrica	24
3.5.1. Control de calidad.	25
3.5.2. Estimación por captura y recaptura.	27
4. Mecánica Estadística	29
4.1. Algunas distribuciones relacionadas con la estadística de Maxwell-Boltzmann	31
4.1.1. Cantidad de partículas por celda: la distribución binomial	31
4.1.2. Forma límite: la distribución de Poisson	32
4.2. Algunas distribuciones relacionadas con la estadística de Bose-Einstein	33
4.2.1. Cantidad de partículas por celda	33
4.2.2. Forma límite: la distribución de Geométrica	34
4.3. Tiempos de espera	35
5. Bibliografía consultada	36

1. Teoría general

1.1. Los axiomas de Kolmogorov

Sean Ω un conjunto no vacío cuyos elementos ω serán llamados *eventos elementales* y \mathcal{A} una familia de subconjuntos de Ω que serán llamados *eventos*.

Definición 1.1. \mathcal{A} es un *álgebra de eventos* si contiene a Ω y es cerrada por complementos y uniones finitas¹

- (i) $\Omega \in \mathcal{A}$,
- (ii) $A \in \mathcal{A}$ implica $A^c \in \mathcal{A}$,
- (iii) $A, B \in \mathcal{A}$ implica $A \cup B \in \mathcal{A}$.

Definición 1.2. Una *medida de probabilidad* \mathbb{P} sobre (Ω, \mathcal{A}) es una función $\mathbb{P} : \mathcal{A} \rightarrow \mathbb{R}$ que satisface los axiomas siguientes:

I. Para cada $A \in \mathcal{A}$, $\mathbb{P}(A) \geq 0$,

II. $\mathbb{P}(\Omega) = 1$.

III. *Aditividad.* Si los eventos A y B no tienen elementos en común, entonces

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

IV. *Axioma de continuidad.* Para cada sucesión decreciente de eventos

$$A_1 \supset A_2 \supset \cdots \supset A_n \supset \cdots, \quad (1)$$

tal que

$$\bigcap_{n=1}^{\infty} A_n = \emptyset$$

vale que

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 0.$$

Definición 1.3. Un *espacio de probabilidad* es una terna $(\Omega, \mathcal{A}, \mathbb{P})$ formada por un conjunto no vacío Ω , llamado *el espacio muestral*; un álgebra \mathcal{A} de subconjuntos de Ω ; llamados *los eventos aleatorios*; y una medida de probabilidad \mathbb{P} definida sobre los eventos aleatorios.

¹**Nomenclatura y definiciones previas.** Sean A y B eventos.

1. Escribiremos $A^c := \{\omega \in \Omega : \omega \notin A\}$ para designar al evento que no ocurre A . El evento A^c se llama el *complemento* de A .
2. Escribiremos $A \cup B := \{\omega \in \Omega : \omega \in A \text{ o } \omega \in B\}$ para designar al evento que ocurre al menos uno de los eventos A o B . El evento $A \cup B$ se llama la *unión* de A y B .
3. Escribiremos $A \cap B := \{\omega \in \Omega : \omega \in A \text{ y } \omega \in B\}$ para designar al evento ocurren ambos A y B . El evento $A \cap B$ se llama la *intersección* de A y B .

A veces escribiremos $A \setminus B$ en lugar de $A \cap B^c$, esto es, el evento que A ocurre, pero B no lo hace. Cuando dos eventos A y B no tienen elementos en común, esto es $A \cap B = \emptyset$, diremos que A y B son *disjuntos*. Una colección de eventos A_1, A_2, \dots se dice *disjunta dos a dos*, si $A_i \cap A_j = \emptyset$ para todo $i \neq j$.

Nota Bene (Consistencia). El sistema de axiomas I-IV es *consistente*. Esto se prueba mediante un ejemplo. Sea Ω un conjunto que consiste de un solo elemento y sea $\mathcal{A} = \{\emptyset, \Omega\}$ la familia de todos los subconjuntos de Ω . \mathcal{A} es un álgebra y la función $\mathbb{P} : \mathcal{A} \rightarrow \mathbb{R}$ definida por $\mathbb{P}(\Omega) := 1$ y $\mathbb{P}(\emptyset) := 0$ es una medida de probabilidad. \square

Construcción de espacios de probabilidad finitos. Los espacios de probabilidad más simples se construyen de la siguiente manera. Se considera un conjunto finito Ω y una función $p : \Omega \rightarrow [0, 1]$ tal que

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

La función p se llama *función de probabilidad* y los números $p(\omega)$, $\omega \in \Omega$, se llaman las *probabilidades de los eventos elementales* $\omega \in \Omega$ o simplemente las *probabilidades elementales*.

El álgebra de eventos, \mathcal{A} , se toma como el conjunto de todos los subconjuntos de Ω y para cada $A \in \mathcal{A}$ se define

$$\mathbb{P}(A) := \sum_{\omega \in A} p(\omega),$$

donde la suma vacía se define como 0.

Todos los espacios de probabilidad finitos en los que \mathcal{A} es la familia de todos los subconjuntos de Ω se construyen de esta manera.

Ejemplo 1.4 (Lanzar una moneda equilibrada). Se lanza una moneda. Los resultados posibles son cara o ceca y pueden representarse mediante las letras *H* (*head*) y *T* (*tail*). Adoptando esa representación el espacio muestral correspondiente es

$$\Omega = \{H, T\}.$$

Decir que una moneda es equilibrada significa que la función de probabilidad asigna igual probabilidad a los dos resultados posibles:

$$p(H) = p(T) = 1/2.$$

\square

Equiprobabilidad: fórmula de Laplace. Sea Ω un espacio muestral finito. Cuando todos los eventos elementales tienen la misma probabilidad, esto es, cuando para todo $\omega \in \Omega$ vale que $p(\omega) = |\Omega|^{-1}$, se dice que el espacio es *equiprobable*. En ese caso las probabilidades de los eventos $A \subset \Omega$ se calculan usando la *fórmula de Laplace*:

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}.$$

En este contexto el problema principal del cálculo de probabilidades consiste determinar la cantidad de eventos elementales favorables a cada evento posible (sin tener que enumerarlo). En otras palabras, la teoría de probabilidades se reduce al *análisis combinatorio*, una importante (y a veces muy difícil) rama de la matemática dedicada a lo que podría llamarse “contar sin contar”. En la Sección 3 se desarrollan sus elementos básicos. \square

1.2. Relación con los datos experimentales

En el mundo real de los experimentos la teoría de probabilidad se aplica de la siguiente manera:

(1) Consideramos un sistema de condiciones, \mathcal{S} , que se pueden repetir cualquier cantidad de veces.

(2) Estudiamos una familia determinada de eventos que pueden ocurrir como resultado de realizar las condiciones \mathcal{S} . En los casos individuales donde se realizan las condiciones \mathcal{S} , los eventos ocurren, generalmente, de distintas maneras. En el conjunto Ω incluimos, *a priori*, todos los resultados que podrían obtenerse al realizar las condiciones \mathcal{S} .

(3) Si al realizar las condiciones \mathcal{S} el resultado pertenece al conjunto A (definido de alguna manera), diremos que ocurre el evento A .

Ejemplo 1.5 (Dos monedas). Las condiciones \mathcal{S} consisten en lanzar una moneda dos veces. El conjunto de los eventos mencionados en (2) resultan del hecho de que en cada lanzamiento puede obtenerse una cara (H) o una ceca (T). Hay cuatro resultados posibles (los eventos elementales), a saber: HH , HT , TH , TT . Si el evento A se define por la ocurrencia de una repetición, entonces A consistirá en que suceda el primero o el cuarto de los cuatro eventos elementales. Esto es, $A = \{HH, TT\}$. De la misma manera todo evento puede considerarse como un conjunto de eventos elementales. \square

(4) Bajo ciertas condiciones se puede suponer que, dado el sistema de condiciones \mathcal{S} , un evento A que a veces ocurre y a veces no, tiene asignado un número real $\mathbb{P}(A)$ que tiene las siguientes características:

(a) Se puede estar prácticamente seguro de que si el sistema de condiciones \mathcal{S} se repite una gran cantidad de veces, n , entonces si $n(A)$ es la cantidad de veces que ocurre el evento A , la proporción $n(A)/n$ diferirá muy poco de $\mathbb{P}(A)$.

(b) Si $\mathbb{P}(A)$ es muy pequeña, se puede estar prácticamente seguro de que cuando se realicen las condiciones \mathcal{S} solo una vez, el evento A no ocurrirá.

Deducción empírica de los axiomas I, II, III. En general, se puede suponer que la familia \mathcal{A} de los eventos observados A, B, C, \dots que tienen probabilidades asignadas, constituye un álgebra de eventos. Está claro que $0 \leq n(A)/n \leq 1$ de modo que el axioma I es bastante natural. Para el evento Ω , $n(\Omega)$ siempre es igual a n de modo que es natural definir $\mathbb{P}(\Omega) = 1$ (Axioma II). Si finalmente, A y B son incompatibles (i.e., no tienen elementos en común), entonces $n(A \cup B) = n(A) + n(B)$ y de aquí resulta que

$$\frac{n(A \cup B)}{n} = \frac{n(A)}{n} + \frac{n(B)}{n}.$$

Por lo tanto, es apropiado postular que $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ (Axioma III).

Nota Bene 1. La afirmación de que un evento A ocurre en las condiciones \mathcal{S} con una determinada probabilidad $\mathbb{P}(A)$ equivale a decir que en una serie suficientemente larga de experimentos (es decir, de realizaciones del sistema de condiciones \mathcal{S}), las frecuencias relativas

$$\hat{p}_k(A) := \frac{n_k(A)}{n_k}$$

de ocurrencia del evento A (donde n_k es la cantidad de experimentos realizados en la k -ésima serie y $n_k(A)$ la cantidad de ellos en los que ocurre A) son aproximadamente idénticas unas a otras y están próximas a $\mathbb{P}(A)$. \square

Ejemplo 1.6. Las condiciones \mathcal{S} consisten en lanzar una moneda (posiblemente cargada). Podemos poner $\Omega = \{H, T\}$ y $\mathcal{A} = \{\emptyset, \{H\}, \{T\}, \Omega\}$, y las posibles medidas de probabilidad $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$ están dadas por

$$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(H) = p, \quad \mathbb{P}(T) = 1 - p, \quad \mathbb{P}(\Omega) = 1,$$

donde p es un número real fijo perteneciente al intervalo $[0, 1]$.

Si en 10 series, de 1000 lanzamientos cada una, se obtienen las siguientes frecuencias relativas de ocurrencia del evento $A = \{H\}$

$$0.753; 0.757; 0.756; 0.750; 0.746; 0.758; 0.751; 0.748; 0.749; 0.746,$$

parece razonable asignarle a p el valor 0.75. \square

Nota Bene 2. Si cada una de dos afirmaciones diferentes es prácticamente segura, entonces podemos decir que simultáneamente son ambas seguras, aunque el grado de seguridad haya disminuido un poco. Si, en cambio, el número de tales afirmaciones es muy grande, de la seguridad práctica de cada una, no podemos deducir nada sobre la validez simultánea de todos ellas. En consecuencia, del principio enunciado en (a) no se deduce que en una cantidad muy grande de series de n experimentos cada una, en *cada uno de ellos* la proporción $n(A)/n$ diferirá sólo un poco de $\mathbb{P}(A)$.

En los casos más típicos de la teoría de probabilidades, la situación es tal que en una larga serie de pruebas es posible obtener uno de los dos valores extremos para la frecuencia

$$\frac{n(A)}{n} = \frac{n}{n} = 1 \quad \text{y} \quad \frac{n(A)}{n} = \frac{0}{n} = 0.$$

Así, cualquiera sea el número de ensayos n , es imposible asegurar con absoluta certeza que tendremos, por ejemplo, la desigualdad

$$\left| \frac{n(A)}{n} - \mathbb{P}(A) \right| < \frac{1}{10}.$$

Por ejemplo, si el evento A es sacar un seis tirando un dado equilibrado, entonces en n tiradas del dado la probabilidad de obtener un seis en todas ellas es $(1/6)^n > 0$; en otras palabras, con probabilidad $(1/6)^n$ tendremos una frecuencia relativa igual a *uno* de sacar un seis en todas las tiradas ; y con probabilidad $(5/6)^n$ no saldrá ningún seis, es decir, la frecuencia relativa de sacar seis será igual a *cero*. \square

Nota Bene 3. De acuerdo con nuestros axiomas a un evento imposible (un conjunto vacío) le corresponde la probabilidad $\mathbb{P}(\emptyset) = 0$, pero la recíproca no es cierta: $\mathbb{P}(A) = 0$ no implica la imposibilidad de A . Cuando $\mathbb{P}(A) = 0$, del principio (b) todo lo que podemos asegurar es que cuando se realicen las condiciones \mathcal{S} una sola vez, el evento A será prácticamente imposible. Sin embargo, esto no asegura de ningún modo que en una sucesión suficientemente grande de experimentos el evento A no ocurrirá. Por otra parte, del principio (a) solamente se puede deducir que cuando $\mathbb{P}(A) = 0$ y n es muy grande, la proporción $n(A)/n$ debe ser muy pequeña (por ejemplo, $1/n$). \square

1.3. Corolarios inmediatos de los axiomas

De $A \cup A^c = \Omega$ y los axiomas II y III se deduce que

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

En particular, debido a que $\Omega^c = \emptyset$, tenemos que $\mathbb{P}(\emptyset) = 0$.

Teorema de aditividad. Si los eventos A_1, A_2, \dots, A_n son disjuntos dos a dos, entonces del axioma III se deduce la fórmula

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

Ejercicios adicionales

1. Sean A y B dos eventos. Mostrar que

(a) Si $A \subseteq B$, entonces $\mathbb{P}(A) \leq \mathbb{P}(B)$. Más precisamente: $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$.

Sugerencia. Expresar el evento B como la unión disjunta de los eventos A y $B \setminus A$ y usar el axioma III.

(b) La probabilidad de que ocurra al menos uno de los eventos A o B es

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

Sugerencia. La unión $A \cup B$ de dos eventos puede expresarse como la unión de dos eventos disjuntos: $A \cup (B \setminus (A \cap B))$.

2. Mostrar que para eventos A , B y C vale que

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

3. Mostrar que para eventos A_1, A_2, \dots, A_n vale que

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_i \mathbb{P}(A_i) - \sum_{i < j} \mathbb{P}(A_i \cap A_j) + \sum_{i < j < k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots \\ &\quad + (-1)^n \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n). \end{aligned}$$

1.4. Sobre el axioma de continuidad

Nota Bene 1. Si la familia de eventos \mathcal{A} es finita el axioma de continuidad IV se deduce de los axiomas I-III. En tal caso, en la sucesión (1) solo hay una cantidad finita de eventos diferentes. Si A_k es el menor de ellos, entonces todos los conjuntos A_{k+m} , $m \geq 1$ coinciden con A_k . Tenemos que $A_k = A_{k+m} = \bigcap_{n=1}^{\infty} A_n = \emptyset$ y $\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}(\emptyset) = 0$. Por lo tanto, todos los ejemplos de espacios de probabilidad *finitos* satisfacen los axiomas I-IV. \square

Nota Bene 2. Se puede probar que para espacios muestrales infinitos, el axioma de continuidad IV es independiente de los axiomas I-III. Este axioma es esencial solamente para espacios de probabilidad infinitos y es casi imposible elucidar su significado empírico en la forma en que lo hicimos con los axiomas I-III. \square

Ejemplo 1.7. Sean $\Omega = \mathbb{Q} \cap [0, 1] = \{r_1, r_2, r_3, \dots\}$ y \mathcal{A}_0 la familia de los subconjuntos de Ω de la forma $[a, b]$, $[a, b)$, $(a, b]$ o (a, b) . La familia, \mathcal{A} de todas las uniones finitas de conjuntos disjuntos de \mathcal{A}_0 es un álgebra de eventos. La medida de probabilidad definida por

$$\begin{aligned}\mathbb{P}(A) &:= b - a, & \text{si } A \in \mathcal{A}_0, \\ \mathbb{P}(A) &:= \sum_{i=1}^k \mathbb{P}(A_i) & \text{si } A = \bigcup_{i=1}^k A_i, \text{ para } A_i \in \mathcal{A}_0 \text{ y } A_i \cap A_j = \emptyset,\end{aligned}$$

satisface los axiomas I-III pero no satisface el axioma de continuidad.

En efecto, para cada $r \in \Omega$, $\{r\} \in \mathcal{A}$ y $\mathbb{P}(\{r\}) = 0$. Los eventos $A_n := \Omega \setminus \{r_1, \dots, r_n\}$, $n \in \mathbb{N}$, son decrecientes y $\bigcap_{n=1}^{\infty} A_n = \emptyset$, sin embargo $\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 1$, debido a que $\mathbb{P}(A_n) = 1$ para todo $n \geq 1$. \square

Teorema 1.8.

- (a) Si $A_1 \supset A_2 \supset \dots$ y $A = \bigcap_{n=1}^{\infty} A_n$, entonces $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$.
- (b) Si $A_1 \subset A_2 \subset \dots$ y $A = \bigcup_{n=1}^{\infty} A_n$, entonces $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$.

Demostración.

(a) Considerar la sucesión $B_n = A_n \setminus A$. Observar que $B_1 \supset B_2 \supset \dots$ y $\bigcap_{n=1}^{\infty} B_n = \emptyset$. Por el axioma de continuidad se obtiene $\lim_{n \rightarrow \infty} \mathbb{P}(B_n) = 0$. Como $\mathbb{P}(B_n) = \mathbb{P}(A_n) - \mathbb{P}(A)$ se deduce que

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}(A).$$

(b) Considerar la sucesión $B_n = A_n^c$. Observar que $B_1 \supset B_2 \supset \dots$ y $\bigcap_{n=1}^{\infty} B_n = A^c$. Por el inciso (a) se obtiene $\lim_{n \rightarrow \infty} \mathbb{P}(B_n) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. Como $\mathbb{P}(B_n) = 1 - \mathbb{P}(A_n)$ se deduce que

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}(A).$$

\square

Ejemplo 1.9 (Números aleatorios). Teóricamente, los números aleatorios son realizaciones independientes del experimento conceptual que consiste en “elegir al azar” un número U del intervalo $(0, 1]$. Aquí la expresión “elegir al azar” significa que el número U tiene la distribución uniforme sobre el intervalo $(0, 1]$, i.e., la probabilidad del evento $U \in (a, b]$ es igual a $b - a$, para cualquier pareja de números reales a y b tales que $0 < a < b \leq 1$. \square

Ejemplo 1.10 (Ternario de Cantor). Se elige al azar un número U del intervalo $(0, 1]$, ¿cuál es la probabilidad de que el 1 no aparezca en el desarrollo en base 3 de U ?

Consideramos la representación en base 3 del número U :

$$U = \sum_{k \geq 1} \frac{a_k(U)}{3^k},$$

donde $a_k(U) \in \{0, 1, 2\}$, $k \geq 1$.

Lo que queremos calcular es la probabilidad del evento $A = \{a_k(U) \neq 1, \forall k \geq 1\}$. Primero observamos que

$$A = \bigcap_{n=1}^{\infty} A_n,$$

donde $A_n = \{a_k(U) \neq 1, \forall 1 \leq k \leq n\}$ y notamos que $A_1 \supset A_2 \supset \dots$. Usando el inciso (a) del **Teorema 1.8** tenemos que $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$. El problema se reduce a calcular la sucesión de probabilidades $\mathbb{P}(A_n)$ y su límite.

Geométricamente el evento A_1 se obtiene eliminando el segmento $(1/3, 2/3)$ del intervalo $(0, 1]$:

$$A_1 = (0, 1/3] \cup [2/3, 1].$$

Para obtener A_2 eliminamos los tercios centrales de los dos intervalos que componen A_1 :

$$A_2 = (0, 1/9] \cup [2/9, 3/9] \cup [6/9, 7/9] \cup [8/9, 1].$$

Continuando de este modo obtenemos una caracterización geométrica de los eventos A_n : A_n es la unión disjunta de 2^n intervalos, cada uno de longitud 3^{-n} . En consecuencia,

$$\mathbb{P}(A_n) = 2^n \frac{1}{3^n} = \left(\frac{2}{3}\right)^n$$

Por lo tanto, $\mathbb{P}(A) = \lim_{n \rightarrow \infty} (2/3)^n = 0$. □

Teorema 1.11 (σ -aditividad). Si A_1, A_2, \dots , es una sucesión de eventos disjuntos dos a dos (i.e., $A_i \cap A_j = \emptyset$ para todos los pares i, j tales que $i \neq j$) y $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$, entonces

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) \tag{2}$$

Demostración. La sucesión de eventos $R_n := \bigcup_{m>n} A_m$, $n \geq 1$, es decreciente y tal que $\bigcap_{n=1}^{\infty} R_n = \emptyset$. Por el axioma IV tenemos que

$$\lim_{n \rightarrow \infty} \mathbb{P}(R_n) = 0 \tag{3}$$

y por el teorema de aditividad tenemos que

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{k=1}^n \mathbb{P}(A_k) + \mathbb{P}(R_n). \tag{4}$$

De (4) y (3) se obtiene (2). □

Corolario 1.12 (Teorema de cubrimiento). Si B, A_1, A_2, \dots es una sucesión de eventos tal que $A = \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$ y $B \subset A$, entonces

$$\mathbb{P}(B) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Demostración. Una cuenta. Descomponemos B en una unión disjunta de eventos

$$B = B \cap \left(\bigcup_{n=1}^{\infty} A_n \right) = \bigcup_{n=1}^{\infty} \left(B \cap \left(A_n \setminus \bigcup_{k=1}^{n-1} (A_n \cap A_k) \right) \right)$$

y aplicamos el teorema de σ -aditividad

$$\mathbb{P}(B) = \sum_{n=1}^{\infty} \mathbb{P} \left(B \cap \left(A_n \setminus \bigcup_{k=1}^{n-1} (A_n \cap A_k) \right) \right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

□

Ejercicios adicionales

4. Sean Ω un conjunto no vacío y \mathcal{A} un álgebra de eventos. Sea $\mathbb{P} : \mathcal{A} \rightarrow \mathbb{R}$ una función tal que

- I. Para cada $A \in \mathcal{A}$, $\mathbb{P}(A) \geq 0$,
- II. $\mathbb{P}(\Omega) = 1$.
- III. Si los eventos A y B no tienen elementos en común, entonces $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
- IV'. Si $(A_n)_{n \geq 1}$ es una sucesión de eventos disjuntos dos a dos y $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$, entonces

$$\mathbb{P} \left(\bigcup_{n=1}^{\infty} A_n \right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Mostrar que bajo esas condiciones la función \mathbb{P} satisface el axioma de continuidad.

1.5. σ -álgebras y teorema de extensión

El álgebra \mathcal{A} se llama una σ -álgebra, si toda unión numerable $\bigcup_{n=1}^{\infty} A_n$ de conjuntos $A_1, A_2, \dots \in \mathcal{A}$, disjuntos dos a dos, también pertenece a \mathcal{A} .

De la identidad

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} \left(A_n \setminus \bigcup_{k=1}^{n-1} (A_n \cap A_k) \right)$$

se deduce que la σ -álgebra también contiene todas las uniones numerables de conjuntos $A_1, A_2, \dots \in \mathcal{A}$. De la identidad

$$\bigcap_{n=1}^{\infty} A_n = \Omega \setminus \bigcup_{n=1}^{\infty} A_n^c$$

lo mismo puede decirse de las intersecciones.

Nota Bene. Solamente cuando disponemos de una medida de probabilidad, \mathbb{P} , definida sobre una σ -álgebra, \mathcal{A} , obtenemos libertad de acción total, sin peligro de que ocurran eventos que no tienen probabilidad.

Lema 1.13 (σ -álgebra generada). Dada un álgebra \mathcal{A} existe la menor σ -álgebra, $\sigma(\mathcal{A})$, que la contiene, llamada la σ -álgebra generada por \mathcal{A} .

Teorema 1.14 (Extensión). Dada una función de conjuntos, \mathbb{P} , no negativa y σ -aditiva definida sobre un álgebra \mathcal{A} se la puede extender a todos los conjuntos de la σ -álgebra generada por \mathcal{A} , $\sigma(\mathcal{A})$, sin perder ninguna de sus propiedades (no negatividad y σ -aditividad) y esta extensión puede hacerse de una sola manera.

Esbozo de la demostración. Para cada $A \subset \Omega$ definimos

$$\mathbb{P}^*(A) := \inf_{A \subset \bigcup_n A_n} \sum_n \mathbb{P}(A_n),$$

donde el ínfimo se toma respecto a todos los cubrimientos del conjunto A por colecciones finitas o numerables de conjuntos A_n pertenecientes a \mathcal{A} . De acuerdo con el Teorema de cubrimiento $\mathbb{P}^*(A)$ coincide con $\mathbb{P}(A)$ para todo conjunto $A \in \mathcal{A}$.

La función \mathbb{P}^* es no negativa y σ -aditiva sobre $\sigma(\mathcal{A})$. La unicidad de la extensión se deduce de la propiedad minimal de $\sigma(\mathcal{A})$. \square

2. Simulación de experimentos aleatorios con espacio muestral finito

2.1. Números aleatorios.

Toda computadora tiene instalado un algoritmo para simular números aleatorios que se pueden obtener mediante una instrucción del tipo “random”. En el *software* Octave, por ejemplo, la sentencia *rand* simula un número aleatorio y *rand(1, n)* simula un vector de n números aleatorios. En algunas calculadoras (llamadas científicas) la instrucción *Ran#* permite simular números aleatorios de tres dígitos. En algunos libros de texto se pueden encontrar tablas de números aleatorios (p. ej., Meyer, P. L.: *Introductory Probability and Statistical Applications*. Addison-Wesley, Massachusetts. (1972))

Cómo usar los números aleatorios. La idea principal se puede presentar mediante un ejemplo muy simple. Queremos construir un mecanismo aleatorio para simular el lanzamiento de una moneda cargada con probabilidad p de obtener de obtener “cara”. Llamemos X al resultado del lanzamiento: $X \in \{0, 1\}$ con la convención de que “cara” = 1 y “ceca” = 0.

Para construir X usamos un número aleatorio U , uniformemente distribuido sobre el intervalo $[0, 1]$ y definimos

$$X := \mathbf{1}\{1 - p < U \leq 1\}. \quad (5)$$

Es fácil ver X satisface las condiciones requeridas. En efecto,

$$\mathbb{P}(X = 1) = \mathbb{P}(1 - p < U \leq 1) = 1 - (1 - p) = p.$$

La ventaja de la construcción es que se puede implementar casi inmediatamente en una computadora. Por ejemplo, si $p = 1/2$, una rutina en Octave para simular X es la siguiente

Rutina para simular el lanzamiento de una moneda equilibrada

```
U = rand;
if U>1/2
    X=1;
else
    X=0;
end
X
```

Nota Bene. El ejemplo anterior es el prototipo para construir y simular experimentos aleatorios. Con la misma idea podemos construir experimentos aleatorios tan complejos como queramos.

2.2. Simulación de experimentos aleatorios

Supongamos que $\Omega = \{\omega_1, \omega_2, \dots, \omega_m\}$ representa el espacio muestral correspondiente a un experimento aleatorio y que cada evento elemental $\omega_k \in \Omega$ tiene asignada la probabilidad $p(\omega_k) = p_k$. Usando un número aleatorio, U , uniformemente distribuido sobre el intervalo $(0, 1]$, podemos construir un mecanismo aleatorio, X , para simular los resultados del experimento aleatorio considerado. Definimos

$$X = \sum_{k=1}^m k \mathbf{1}\{L_{k-1} < U \leq L_k\}, \quad (6)$$

donde

$$L_0 := 0 \quad \text{y} \quad L_k := \sum_{i=1}^k p_i, \quad (1 \leq k \leq m)$$

e identificamos cada evento elemental $\omega_k \in \Omega$ con su correspondiente subíndice k .

En efecto, de la definición (6) se deduce que para cada $k = 1, \dots, m$ vale que

$$\mathbb{P}(X = k) = \mathbb{P}(L_{k-1} < U \leq L_k) = L_k - L_{k-1} = p_k.$$

□

Nota Bene. El mecanismo aleatorio definido en (6) se puede construir “gráficamente” de la siguiente manera:

1. Partir el intervalo $(0, 1]$ en m subintervalos sucesivos I_1, \dots, I_m de longitudes p_1, \dots, p_m , respectivamente.
2. Sortear un número aleatorio, U , y observar en qué intervalo de la partición cae.
3. Si U cae en el intervalo I_k , producir el resultado ω_k .

Ejemplo 2.1 (Lanzar un dado equilibrado). Se quiere simular el lanzamiento de un dado equilibrado. El espacio muestral es $\Omega = \{1, 2, 3, 4, 5, 6\}$ y la función de probabilidades es $p(k) = 1/6$, $k = 1, \dots, 6$. El mecanismo aleatorio $X = X(U)$, definido en (6), se construye de la siguiente manera:

1. Partir el intervalo $(0, 1]$ en 6 intervalos sucesivos de longitud $1/6$: $I_1 = (0, 1/6]$, $I_2 = (1/6, 2/6]$, $I_3 = (2/6, 3/6]$, $I_4 = (3/6, 4/6]$, $I_5 = (4/6, 5/6]$ e $I_6 = (5/6, 6/6]$.
2. Sortear un número aleatorio U .
3. Si $U \in I_k$, $X = k$.

En pocas palabras,

$$X = \sum_{k=1}^6 k \mathbf{1} \left\{ \frac{k-1}{6} < U \leq \frac{k}{6} \right\}. \quad (7)$$

Por ejemplo, si sorteamos un número aleatorio, U y se obtiene que $U = 0.62346$, entonces el valor simulado del dado es $X = 4$. Una rutina en Octave para simular X es la siguiente

Rutina para simular el lanzamiento de un dado

```
U=rand;
k=0;
do
  k++;
until((k-1)/6<U & U<=k/6)
X=k
```

2.3. Estimación de probabilidades

Formalmente, un experimento aleatorio se describe mediante un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Todas las preguntas asociadas con el experimento pueden reformularse en términos de este espacio. En la práctica, decir que un evento A ocurre con una determinada probabilidad $\mathbb{P}(A) = p$ equivale a decir que en una serie suficientemente grande de experimentos las frecuencias relativas de ocurrencia del evento A

$$\hat{p}_k(A) = \frac{n_k(A)}{n_k}$$

(donde n_k es la cantidad de ensayos realizados en la k -ésima serie y $n_k(A)$ es la cantidad en los que ocurre A) son aproximadamente idénticas unas a otras y están próximas a p . Las series de experimentos se pueden simular en una computadora utilizando un *generador de números aleatorios*.

Ejemplo 2.2. El experimento consiste en lanzar 5 monedas equilibradas y registrar la cantidad N de *caras* observadas. El conjunto de todos los resultados posibles es $\Omega = \{0, 1, 2, 3, 4, 5\}$. El problema consiste en asignarle probabilidades a los eventos elementales.

La solución experimental del problema se obtiene realizando *una serie suficientemente grande de experimentos* y asignando a cada evento elemental su frecuencia relativa.

Sobre la base de una rutina similar a la que presentamos en la sección 2.1 para simular el resultado del lanzamiento de una moneda equilibrada se pueden simular $n = 10000$ realizaciones del experimento que consiste en lanzar 5 monedas equilibradas. Veamos como hacerlo. Usamos la construcción (5) para simular el lanzamiento de 5 monedas equilibradas X_1, X_2, X_3, X_4, X_5 . La cantidad de caras observadas es la suma de las X_i : $N = X_1 + X_2 + X_3 + X_4 + X_5$.

Repetiendo la simulación 10000 veces (o genéricamente n veces), obtenemos una tabla que contiene la cantidad de veces que fué simulado cada valor de la variable N . Supongamos que obtuvimos la siguiente tabla:

valor simulado	0	1	2	3	4	5
cantidad de veces	308	1581	3121	3120	1564	306

(8)

En tal caso diremos que se obtuvieron las siguientes estimaciones

$$\begin{aligned} \mathbb{P}(N = 0) &\approx 0.0308, & \mathbb{P}(N = 1) &\approx 0.1581, & \mathbb{P}(N = 2) &\approx 0.3121, \\ \mathbb{P}(N = 3) &\approx 0.3120, & \mathbb{P}(N = 4) &\approx 0.1564, & \mathbb{P}(N = 5) &\approx 0.0306. \end{aligned}$$

Para finalizar este ejemplo, presentamos un programa en Octave que simula diez mil veces el lanzamiento de cinco monedas equilibradas, contando en cada una la cantidad de caras observadas y que al final provee una tabla como la representada en (8)

```

n = 10000;
N = zeros(1,n);
for i=1:n
    U=rand(1,5);
    X=[U<=(1/2)];
    N(i)=sum(X);
end
for j=1:6
    T(j)=sum([N==j-1]);
end
T

```

Nota Bene. Usando las herramientas que proporciona el análisis combinatorio (ver sección 3) se puede demostrar que para cada $k \in \{0, 1, 2, 3, 4, 5\}$ vale que

$$\mathbb{P}(N = k) = \binom{5}{k} \frac{1}{32}.$$

En otros términos,

$$\begin{aligned}\mathbb{P}(N = 0) &= 0.03125, & \mathbb{P}(N = 1) &= 0.15625, & \mathbb{P}(N = 2) &= 0.31250, \\ \mathbb{P}(N = 3) &= 0.31250, & \mathbb{P}(N = 4) &= 0.15625, & \mathbb{P}(N = 5) &= 0.03125.\end{aligned}$$

□

Ejemplo 2.3 (Paradoja de De Mere). ¿Cuál de las siguientes apuestas es más conveniente?

- Obtener al menos un as en 4 tiros de un dado.
- Obtener al menos un doble as en 24 tiros de dos dados.

1. La construcción (7) permite simular 4 tiros de un dado usando 4 números aleatorios independientes U_1, U_2, U_3, U_4 .

La cantidad de ases obtenidos en los 4 tiros es la suma $S = \sum_{i=1}^4 \mathbf{1}\{0 < U_i \leq 1/6\}$. El evento $A_1 = \text{"obtener al menos un as en 4 tiros de un dado"}$ equivale al evento $S \geq 1$.

Si repetimos la simulación 10000 veces podemos obtener una estimación (puntual) de la probabilidad del evento A_1 calculando su frecuencia relativa.

La siguiente rutina (en Octave) provee una estimación de la probabilidad del evento A_1 basada en la repetición de 10000 simulaciones del experimento que consiste en tirar 4 veces un dado.

Rutina 1

```
n=10000;
A1=zeros(1,n);
for i=1:n
    U=rand(1,4);
    S=sum(U<=1/6);
    if S>=1
        A1(i)=1;
    else
        A1(i)=0;
    end
end
hpA1=sum(A1)/n
```

Ejecutando 10 veces la **Rutina 1** se obtuvieron los siguientes resultados para la frecuencia relativa del evento A_1

0.5179 0.5292 0.5227 0.5168 0.5204 0.5072 0.5141 0.5177 0.5127 0.5244

Notar que los resultados obtenidos se parecen entre sí e indican que la probabilidad de obtener al menos un as en 4 tiros de un dado es mayor que 0.5.

2. La construcción (7) permite simular 24 tiros de dos dados usando 48 números aleatorios independientes $U_1, U_2, \dots, U_{47}, U_{48}$.

La cantidad de veces que se obtiene un doble as en los 24 tiros de dos dados es la suma $S = \sum_{i=1}^{24} \mathbf{1}\{0 < U_{2i-1} \leq 1/6, 0 < U_{2i} \leq 1/6\}$. El evento $A_2 = \text{"obtener al menos un doble as en 24 tiros de dos dados"}$ equivale al evento $S \geq 1$.

Si repetimos la simulación 10000 veces podemos obtener una estimación (puntual) de la probabilidad del evento A_2 calculando su frecuencia relativa.

La siguiente rutina (en Octave) provee una estimación de la probabilidad del evento A_2 basada en la repetición de 10000 simulaciones del experimento que consiste en tirar 24 veces dos dados.

Rutina 2

```
n=10000;
A2=zeros(1,n);
for i=1:n
    U=rand(2,24);
    V=(U<=1/6);
    S=sum(V(1,:).*V(2,:));
    if S>=1
        A2(i)=1;
    else
        A2(i)=0;
    end
end
hpA2=sum(A2)/n
```

Ejecutando 10 veces la **Rutina 2** se obtuvieron los siguientes resultados para la frecuencia relativa del evento A_2

0.4829 0.4938 0.4874 0.4949 0.4939 0.4873 0.4882 0.4909 0.4926 0.4880

Notar que los resultados obtenidos se parecen entre sí e indican que la probabilidad de obtener al menos un doble as en 24 tiros de dos dados es menor que 0.5.

Conclusión. Los resultados experimentales obtenidos indican que es mejor apostar a que se obtiene al menos un as en 4 tiros de un dado que apostar a que se obtiene al menos un doble as en 24 tiros de un dado.

3. Elementos de Análisis Combinatorio

Cuando se estudian juegos de azar, procedimientos muestrales, problemas de orden y ocupación, se trata por lo general con espacios muestrales finitos Ω en los que a todos los eventos elementales se les atribuye igual probabilidad. Para calcular la probabilidad de un evento A tenemos que dividir la cantidad de eventos elementales contenidos en A (llamados *casos favorables*) entre la cantidad de total de eventos elementales contenidos en Ω (llamados *casos posibles*). Estos cálculos se facilitan por el uso sistemático de unas pocas reglas.

3.1. Regla del Producto

Sean A y B dos conjuntos cualesquiera. El producto cartesiano de A y B se define por $A \times B = \{(a, b) : a \in A \text{ y } b \in B\}$. Si A y B son finitos, entonces $|A \times B| = |A| \cdot |B|$.

Demostración. Supongamos que $A = \{a_1, a_2, \dots, a_m\}$ y $B = \{b_1, b_2, \dots, b_n\}$. Basta observar el cuadro siguiente

	b_1	b_2	\dots	b_n
a_1	(a_1, b_1)	(a_1, b_2)	\dots	(a_1, b_n)
a_2	(a_2, b_1)	(a_2, b_2)	\dots	(a_2, b_n)
\vdots	\vdots	\vdots		\vdots
a_m	(a_m, b_1)	(a_m, b_2)	\dots	(a_m, b_n)

Cuadro 1: Esquema rectangular del tipo *tabla de multiplicar* con m filas y n columnas: en la intersección de fila i y la columna j se encuentra el par (a_i, b_j) . Cada par aparece una y sólo una vez.

En palabras, con m elementos a_1, \dots, a_m y n elementos b_1, \dots, b_n es posible formar $m \cdot n$ pares (a_i, b_j) que contienen un elemento de cada grupo. \square

Teorema 3.1 (Regla del producto). Sean A_1, A_2, \dots, A_n , n conjuntos cualesquiera. El producto cartesiano de los n conjuntos A_1, A_2, \dots, A_n se define por

$$A_1 \times A_2 \times \dots \times A_n = \{(x_1, x_2, \dots, x_n) : x_i \in A_i, 1 \leq i \leq n\}.$$

Si los conjuntos A_1, A_2, \dots, A_n son finitos, entonces

$$|A_1 \times A_2 \times \dots \times A_n| = \prod_{i=1}^n |A_i|.$$

Demostración. Si $n = 2$ ya lo demostramos. Si $n = 3$, tomamos los pares (x_1, x_2) como elementos de un nuevo tipo. Hay $|A_1| \cdot |A_2|$ elementos de ese tipo y $|A_3|$ elementos x_3 . Cada terna (x_1, x_2, x_3) es un par formado por un elemento (x_1, x_2) y un elemento x_3 ; por lo tanto, la cantidad de ternas es $|A_1| \cdot |A_2| \cdot |A_3|$. Etcétera. \square

Nota Bene. Muchas aplicaciones se basan en la siguiente reformulación de la regla del producto: *r decisiones sucesivas con exactamente n_k elecciones posibles en el k-ésimo paso pueden producir un total de $n_1 \cdot n_2 \cdots n_r$ resultados diferentes.* \square

Ejemplo 3.2 (Ubicar r bolas en n urnas). Los resultados posibles del experimento se pueden representar mediante el conjunto

$$\Omega = \{1, 2, \dots, n\}^r = \{(x_1, x_2, \dots, x_r) : x_i \in \{1, 2, \dots, n\}, 1 \leq i \leq r\},$$

donde $x_i = j$ representa el resultado “la bola i se ubicó en la urna j ”. Cada bola puede ubicarse en una de las n urnas posibles. Con r bolas tenemos r elecciones sucesivas con exactamente n elecciones posibles en cada paso. En consecuencia, r bolas pueden ubicarse en n urnas de n^r formas distintas.

Usamos el lenguaje figurado de bolas y urnas, pero el mismo espacio muestral admite muchas interpretaciones distintas. Para ilustrar el asunto *listaremos una cantidad de situaciones en las cuales aunque el contenido intuitivo varía son todas abstractamente equivalentes al esquema de ubicar r bolas en n urnas, en el sentido de que los resultados difieren solamente en su descripción verbal.*

1. *Nacimientos.* Las configuraciones posibles de los nacimientos de r personas corresponde a los diferentes arreglos de r bolas en $n = 365$ urnas (suponiendo que el año tiene 365 días).
2. *Accidentes.* Clasificar r accidentes de acuerdo con el día de la semana en que ocurrieron es equivalente a poner r bolas en $n = 7$ urnas.
3. *Muestreo.* Un grupo de personas se clasifica de acuerdo con, digamos, edad o profesión. Las clases juegan el rol de las urnas y las personas el de las bolas.
4. *Dados.* Los posibles resultados de una tirada de r dados corresponde a poner r bolas en $n = 6$ urnas. Si en lugar de dados se lanzan monedas tenemos solamente $n = 2$ urnas.
5. *Dígitos aleatorios.* Los posibles ordenamientos de una sucesión de r dígitos corresponden a las distribuciones de r bolas (= lugares) en diez urnas llamadas $0, 1, \dots, 9$.
6. *Coleccionando figuritas.* Los diferentes tipos de figuritas representan las urnas, las figuritas colecciónadas representan las bolas.

□

3.2. Muestras ordenadas

Se considera una “población” de n elementos a_1, a_2, \dots, a_n . Cualquier secuencia ordenada $a_{j_1}, a_{j_2}, \dots, a_{j_k}$ de k símbolos se llama una *muestra ordenada de tamaño k* tomada de la población. (Intuitivamente los elementos se pueden elegir uno por uno). Hay dos procedimientos posibles.

(a) Muestreo con reposición. Cada elección se hace entre toda la población, por lo que cada elemento se puede elegir más de una vez. Cada uno de los k elementos se puede elegir en n formas: la cantidad de muestras posibles es, por lo tanto, n^k , lo que resulta de la regla del producto con $n_1 = n_2 = \dots = n_k = n$.

(b) Muestreo sin reposición. Una vez elegido, el elemento se quita de la población, de modo que las muestras son arreglos sin repeticiones. El volumen de la muestra k no puede exceder el tamaño de la población total n .

Tenemos n elecciones posibles para el primer elemento, pero sólo $n - 1$ para el segundo, $n - 2$ para el tercero, etcétera. Usando la regla del producto se obtiene un total de

$$(n)_k := n(n - 1)(n - 2) \cdots (n - k + 1) \quad (9)$$

elecciones posibles.

Teorema 3.3. Para una población de n elementos y un tamaño de muestra prefijado k , existen n^k diferentes muestras con reposición y $(n)_k$ muestras sin reposición.

Ejemplo 3.4. Consideramos una urna con 8 bolas numeradas $1, 2, \dots, 8$

- (a) **Extracción con reposición.** Extraemos 3 bolas *con reposición*: después de extraer una bola, anotamos su número y la ponemos de nuevo en la urna. El espacio muestral Ω_1 correspondiente a este experimento consiste de todas las secuencias de longitud 3 que pueden formarse con los símbolos $1, 2, \dots, 8$. De acuerdo con el Teorema 3.3, Ω_1 tiene $8^3 = 512$ elementos. Bajo la hipótesis de que todos los elementos tienen la misma probabilidad, la probabilidad de observar la secuencia $(3, 7, 1)$ es $1/512$.
- (b) **Extracción de una colección ordenada sin reposición.** Extraemos 3 bolas *sin reposición*: cada bola elegida *no* se vuelve a poner en la urna. Anotamos los números de las bolas en el orden en que fueron extraídas de la urna. El espacio muestral Ω_2 correspondiente a este experimento es el conjunto de todas las secuencias de longitud 3 que pueden formarse con los símbolos $1, 2, \dots, 8$ donde cada símbolo puede aparecer a lo sumo una vez. De acuerdo con el Teorema 3.3, Ω_2 tiene $(8)_3 = 8 \cdot 7 \cdot 6 = 336$ elementos. Bajo la hipótesis que todos los elementos tienen la misma probabilidad, la probabilidad de observar la secuencia $(3, 7, 1)$ (en ese orden) es $1/336$.

□

Ejemplo 3.5. Una urna contiene 6 bolas rojas y 4 bolas negras. Se extraen 2 bolas con reposición. Para fijar ideas supongamos que las bolas están numeradas de la siguiente manera: las primeras 6 son las rojas y las últimas 4 son las negras. El espacio muestral asociado es $\Omega = \{1, \dots, 10\}^2$ y su cantidad de elementos $|\Omega| = 10^2$.

- (a) ¿Cuál es la probabilidad de que las dos sean rojas? Sea R el evento “las dos son rojas”, $R = \{1, \dots, 6\}^2$ y $|R| = 6^2$. Por lo tanto, $\mathbb{P}(R) = 6^2/10^2 = 0.36$.
- (b) ¿Cuál es la probabilidad de que las dos sean del mismo color? Sea N el evento “las dos son negras”, $N = \{7, \dots, 10\}^2$ y $|N| = 4^2$, entonces $\mathbb{P}(N) = 4^2/10^2 = 0.16$. Por lo tanto, $\mathbb{P}(R \cup N) = \mathbb{P}(R) + \mathbb{P}(N) = 0.52$.
- (c) ¿Cuál es la probabilidad de que al menos una de las dos sea roja? El evento “al menos una de las dos es roja” es el complemento de “las dos son negras”. Por lo tanto, $\mathbb{P}(N^c) = 1 - \mathbb{P}(N) = 0.84$.

Si se consideran extracciones sin reposición, deben reemplazarse las cantidades $(10)^2, 6^2$ y 4^2 por las correspondientes $(10)_2, (6)_2$ y $(4)_2$. □

Caso especial $k = n$. En muestreo sin reposición una muestra de tamaño n incluye a toda la población y representa una *permutación* de sus elementos. En consecuencia, n elementos a_1, a_2, \dots, a_n se pueden ordenar de $(n)_n = n \cdot (n - 1) \cdots 2 \cdot 1$ formas distintas. Usualmente el número $(n)_n$ se denota $n!$ y se llama el *factorial de n* .

Corolario 3.6. *La cantidad de formas distintas en que se pueden ordenar n elementos es*

$$n! = 1 \cdot 2 \cdots n. \quad (10)$$

Observación 3.7. Las muestras ordenadas de tamaño k , sin reposición, de una población de n elementos, se llaman *variaciones* de n elementos tomados de a k . Su número total $(n)_k$ se puede calcular del siguiente modo

$$(n)_k = \frac{n!}{(n - k)!} \quad (11)$$

Nota Bene sobre muestreo aleatorio. Cuando hablamos de “*muestras aleatorias de tamaño k* ”, el adjetivo aleatorio indica que todas las muestras posibles tienen la misma probabilidad, a saber: $1/n^k$ en muestreo con reposición y $1/(n)_k$ en muestreo sin reposición. En ambos casos, n es el tamaño de la población de la que se extraen las muestras.

Si n es grande y k es relativamente pequeño, el cociente $(n)_k/n^k$ está cerca de la unidad. En otras palabras, para grandes poblaciones y muestras relativamente pequeñas, las dos formas de muestrear son prácticamente equivalentes.

Ejemplos

Consideramos muestras aleatorias de volumen k (*con reposición*) tomadas de una población de n elementos a_1, \dots, a_n . Nos interesa el evento que en una muestra no se repita ningún elemento. En total existen n^k muestras diferentes, de las cuales $(n)_k$ satisfacen la condición estipulada. Por lo tanto, *la probabilidad de ninguna repetición en nuestra muestra es*

$$p = \frac{(n)_k}{n^k} = \frac{n(n - 1) \cdots (n - k + 1)}{n^k} \quad (12)$$

Las interpretaciones concretas de la fórmula (12) revelan aspectos sorprendentes.

Muestras aleatorias de números. La población consiste de los diez dígitos $0, 1, \dots, 9$. Toda sucesión de cinco dígitos representa una muestra de tamaño $k = 5$, y supondremos que cada uno de esos arreglos tiene probabilidad 10^{-5} . *La probabilidad de que 5 dígitos aleatorios sean todos distintos es $p = (10)_5 10^{-5} = 0.3024$.*

Bolas y urnas. *Si n bolas se ubican aleatoriamente en n urnas, la probabilidad de que cada urna esté ocupada es*

$$p = \frac{n!}{n^n}.$$

Interpretaciones:

- (a) Para $n = 7$, $p = 0.00612\dots$. Esto significa que *si en una ciudad ocurren 7 accidentes por semana, entonces (suponiendo que todas las ubicaciones posibles son igualmente probables) prácticamente todas las semanas contienen días con dos o más accidentes, y en promedio solo una semana de 164 mostrará una distribución uniforme de un accidente por día.*
- (b) Para $n = 6$ la probabilidad p es igual a $0.01543\dots$. Esto muestra lo extremadamente improbable que en seis tiradas de un dado perfecto aparezcan todas las caras.

Cumpleaños. Los cumpleaños de k personas constituyen una muestra de tamaño k de la población formada por todos los días del año.

De acuerdo con la ecuación (12) la probabilidad, p_k , de que todos los k cumpleaños sean diferentes es

$$p_k = \frac{(365)_k}{365^k} = \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{k-1}{365}\right).$$

Una fórmula aparentemente abominable. Si $k = 23$ tenemos $p_k < 1/2$. En palabras, *para 23 personas la probabilidad que al menos dos personas tengan un cumpleaños común excede 1/2.*

Aproximaciones numéricas de p_k . Si k es chico, tomando logaritmos y usando que para x pequeño y positivo $\log(1 - x) \sim -x$, se obtiene

$$\log p_k \sim -\frac{1 + 2 + \cdots + (k-1)}{365} = -\frac{k(k-1)}{730}.$$

Ejercicios adicionales

- 5.** Hallar la probabilidad p_k de que en una muestra de k dígitos aleatorios no haya dos iguales. Estimar el valor numérico de p_{10} usando la *fórmula de Stirling* (1730): $n! \sim e^{-n} n^{n+\frac{1}{2}} \sqrt{2\pi}$.
- 6.** Considerar los primeros 10000 decimales del número π . Hay 2000 grupos de cinco dígitos. Contar la cantidad de grupos en los que los 5 dígitos son diferentes e indicar la frecuencia relativa del evento considerado. Comparar el resultado obtenido con la probabilidad de que en una muestra de 5 dígitos aleatorios no haya dos iguales.

3.3. Subpoblaciones

En lo que sigue, utilizaremos el término *población de tamaño n* para designar una colección de n elementos *sin considerar su orden*. Dos poblaciones se consideran diferentes si una de ellas contiene algún elemento que no está contenido en la otra.

Uno de los problemas más importantes del cálculo combinatorio es *determinar la cantidad $C_{n,k}$ de subpoblaciones distintas de tamaño k que tiene una población de tamaño n .* Cuando n y k son pequeños, el problema se puede resolver por enumeración directa. Por ejemplo, hay seis formas distintas elegir dos letras entre cuatro letras A, B, C, D , a saber: AB, AC, AD, BC, BD, CD . Así, $C_{4,2} = 6$. Cuando la cantidad de elementos de la colección es grande la enumeración directa es impracticable. El problema general se resuelve razonando

de la siguiente manera: consideramos una subpoblación de tamaño k de una población de n elementos. Cada numeración arbitraria de los elementos de la subpoblación la convierte en una muestra ordenada de tamaño k . Todas las muestras ordenadas de tamaño k se pueden obtener de esta forma. Debido a que k elementos se pueden ordenar de $k!$ formas diferentes, resulta que $k!$ veces la cantidad de subpoblaciones de tamaño k coincide con la cantidad de muestras ordenadas de dicho tamaño. En otros términos, $C_{n,k} \cdot k! = (n)_k$. Por lo tanto,

$$C_{n,k} = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!}. \quad (13)$$

Los números definidos en (13) se llaman *coeficientes binomiales* o *números combinatorios* y la notación clásica para ellos es $\binom{n}{k}$.

Teorema 3.8. *Una población de n elementos tiene*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (14)$$

diferentes subpoblaciones de tamaño $k \leq n$.

Ejemplo 3.9. Consideramos una urna con 8 bolas numeradas 1, 2, ..., 8. Extraemos 3 bolas simultáneamente, de modo que el orden es irrelevante. El espacio muestral Ω_3 correspondiente a este experimento consiste de todos los subconjuntos de tamaño 3 del conjunto $\{1, 2, \dots, 8\}$. Por el Teorema 3.8 Ω_3 tiene $\binom{8}{3} = 56$ elementos. Bajo la hipótesis de que todos los elementos tienen la misma probabilidad, la probabilidad de seleccionar $\{3, 7, 1\}$ es 1/56. \square

Dada una población de tamaño n podemos elegir una subpoblación de tamaño k de $\binom{n}{k}$ maneras distintas. Ahora bien, elegir los k elementos que vamos a quitar de una población es lo mismo que elegir los $n - k$ elementos que vamos a dejar dentro. Por lo tanto, es claro que para cada $k \leq n$ debe valer

$$\binom{n}{k} = \binom{n}{n-k}. \quad (15)$$

La ecuación (15) se deduce inmediatamente de la identidad (14). El lado izquierdo de la ecuación (15) no está definido para $k = 0$, pero el lado derecho si lo está. Para que la ecuación (15) sea válida para todo entero k tal que $0 \leq k \leq n$, se definen

$$\binom{n}{0} := 1, \quad 0! := 1, \quad \text{y} \quad (n)_0 := 1.$$

Triángulo de Pascal. Las ecuaciones en diferencias

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}, \quad (16)$$

junto con el conocimiento de los datos de borde

$$\binom{n}{0} = \binom{n}{n} = 1, \quad (17)$$

determinan completamente los números combinatorios $\binom{n}{k}$, $0 \leq k \leq n$, $n = 0, 1, \dots$. Usando dichas relaciones se construye el famoso “*triángulo de Pascal*”, que muestra todos los números combinatorios en la forma de un triángulo

$$\begin{array}{ccccccc}
& & & 1 & & & \\
& & & 1 & 1 & & \\
& & & 1 & 2 & 1 & \\
& & & 1 & 3 & 3 & 1 \\
& & & 1 & 4 & 6 & 4 & 1 \\
& & & 1 & 5 & 10 & 10 & 5 & 1 \\
1 & 6 & 15 & 20 & 15 & 6 & 1 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots
\end{array}$$

La n -ésima fila de este triángulo contiene los coeficientes $\binom{n}{0}, \binom{n}{1}, \dots, \binom{n}{n}$. Las condiciones de borde (17) indican que el primero y el último de esos números son 1. Los números restantes se determinan por la ecuación en diferencias (16). Vale decir, para cada $0 < k < n$, el k -ésimo coeficiente de la n -ésima fila del “*triángulo de Pascal*” se obtiene *sumando* los dos coeficientes inmediatamente superiores a izquierda y derecha. Por ejemplo, $\binom{5}{2} = 4 + 6 = 10$.

Control de calidad. Una planta de ensamblaje recibe una partida de 50 piezas de precisión que incluye 4 defectuosas. La división de control de calidad elige 10 piezas al azar para controlarlas y rechaza la partida si encuentra 1 o más defectuosas. ¿Cuál es la probabilidad de que la partida pase la inspección? Hay $\binom{50}{10}$ formas de elegir la muestra para controlar y $\binom{46}{10}$ de elegir todas las piezas sin defectos. Por lo tanto, la probabilidad es

$$\binom{46}{10} \binom{50}{10}^{-1} = \frac{46!}{10!36!} \frac{10!40!}{50!} = \frac{40 \cdot 39 \cdot 38 \cdot 37}{50 \cdot 49 \cdot 48 \cdot 47} = 0,3968\dots$$

Usando cálculos casi idénticos una compañía puede decidir sobre qué cantidad de piezas defectuosas admite en una partida y diseñar un programa de control con una probabilidad dada de éxito. \square

Ejercicios adicionales

7. Considerar el siguiente juego: el jugador I tira 4 veces una moneda honesta y el jugador II lo hace 3 veces. Calcular la probabilidad de que el jugador I obtenga más caras que el jugador II.

3.4. Particiones

Teorema 3.10. Sean r_1, \dots, r_k enteros tales que

$$r_1 + r_2 + \dots + r_k = n, \quad r_i \geq 0. \tag{18}$$

El número de formas en que una población de n elementos se puede dividir en k partes ordenadas (particionarse en k subpoblaciones) tales que la primera contenga r_1 elementos, la

segunda r_2 , etc, es

$$\frac{n!}{r_1!r_2!\cdots r_k!}. \quad (19)$$

Los números (19) se llaman coeficientes multinomiales.

Demostración. Un uso repetido de (14) muestra que el número (19) se puede reescribir en la forma

$$\binom{n}{r_1} \binom{n-r_1}{r_2} \binom{n-r_1-r_2}{r_3} \cdots \binom{n-r_1-\cdots-r_{k-2}}{r_{k-1}} \quad (20)$$

Por otro lado, para efectuar la partición deseada, tenemos primero que seleccionar r_1 elementos de los n ; de los restantes $n - r_1$ elementos seleccionamos un segundo grupo de tamaño r_2 , etc. Después de formar el grupo $(k - 1)$ quedan $n - r_1 - r_2 - \cdots - r_{k-1} = r_k$ elementos, y esos forman el último grupo. Concluimos que (20) representa el número de formas en que se puede realizar la partición. \square

Ejemplo 3.11 (Accidentes). En una semana ocurrieron 7 accidentes. Cuál es la probabilidad de que en dos días de esa semana hayan ocurrido dos accidentes cada día y de que en otros tres días hayan ocurrido un accidente cada día?

Primero particionamos los 7 días en 3 subpoblaciones: dos días con dos accidentes en cada uno, tres días con un accidente en cada uno y dos días sin accidentes.. Esa partición en tres grupos de tamaños 2, 3, 2 se puede hacer de $7!/(2!3!2!)$ formas distintas y por cada una de ellas hay $7!/(2!2!1!1!0!0!) = 7!/(2!2!)$ formas diferentes de ubicar los 7 accidentes en los 7 días. Por lo tanto, el valor de la probabilidad requerido es igual a

$$\frac{7!}{2!3!2!} \times \frac{7!}{2!2!} \frac{1}{7^7} = 0.3212\dots$$

\square

Ejercicios adicionales

8. ¿Cuántas palabras distintas pueden formarse permutando las letras de la palabra “manzana” y cuántas permutando las letras de la palabra “aiaiaiaiaiaiii”?
9. Se ubicarán 6 bolas distinguibles en 8 urnas numeradas 1, 2, ..., 8. Suponiendo que todas las configuraciones distintas son equiprobables calcular la probabilidad de que resulten tres urnas ocupadas con una bola cada una y que otra urna contenga las tres bolas restantes.

3.5. Distribución Hipergeométrica

Muchos problemas combinatorios se pueden reducir a la siguiente forma. En una urna hay n_1 bolas rojas y n_2 bolas negras. Se elige al azar un grupo de r bolas. Se quiere calcular la probabilidad p_k de que en el grupo elegido, haya exactamente k bolas rojas, $0 \leq k \leq \min(n_1, r)$.

Para calcular p_k , observamos que el grupo elegido debe contener k bolas rojas y $r-k$ negras. Las rojas pueden elegirse de $\binom{n_1}{k}$ formas distintas y las negras de $\binom{n_2}{r-k}$ formas distintas. Como cada elección de las k bolas rojas debe combinarse con cada elección de las $r-k$ negras, se obtiene

$$p_k = \binom{n_1}{k} \binom{n_2}{r-k} \binom{n_1 + n_2}{r}^{-1} \quad (21)$$

El sistema de probabilidades obtenido se llama la *distribución hipergeométrica*.

3.5.1. Control de calidad.

En control de calidad industrial, se someten a inspección lotes de n unidades. Las unidades defectuosas juegan el rol de las bolas rojas y su cantidad n_1 es desconocida. Se toma una muestra de tamaño r y se determina la cantidad k de unidades defectuosas. La fórmula (21) permite hacer inferencias sobre la cantidad desconocida n_1 ; se trata de problema típico de estimación estadística que será analizado más adelante. \square

Ejemplo 3.12. Una planta de ensamblaje recibe una partida de 100 piezas de precisión que incluye exactamente 8 defectuosas. La división control de calidad elige 10 piezas al azar para controlarlas y rechaza la partida si encuentra al menos 2 defectuosas. ¿Cuál es la probabilidad de que la partida pase la inspección?

El criterio de decisión adoptado indica que la partida pasa la inspección si (y sólo si) en la muestra no se encuentran piezas defectuosas o si se encuentra exactamente una pieza defectuosa. Hay $\binom{100}{10}$ formas de elegir la muestra para controlar, $\binom{92}{0}\binom{8}{0}$ formas de elegir muestras sin piezas defectuosas y $\binom{92}{1}\binom{8}{1}$ formas de elegir muestras con exactamente una pieza defectuosa. En consecuencia la probabilidad de que la partida pase la inspección es

$$\binom{92}{10} \binom{8}{0} \binom{100}{10}^{-1} + \binom{92}{9} \binom{8}{1} \binom{100}{10}^{-1} \approx 0.818.$$

\square

Ejemplo 3.13. Una planta de ensamblaje recibe una partida de 100 piezas de precisión que incluye exactamente k defectuosas. La división control de calidad elige 10 piezas al azar para controlarlas y rechaza la partida si encuentra al menos 2 defectuosas. ¿Con ese criterio de decisión, cómo se comporta la probabilidad $p(k)$ de que la partida pase la inspección?

Una partida pasará la inspección si (y sólo si) al extraer una muestra de control la cantidad de piezas defectuosas encontradas es 0 o 1. Hay $\binom{100}{10}$ formas de elegir la muestra para controlar. Para cada $k = 1, \dots, 90$ hay $\binom{100-k}{10-k} \binom{k}{0}$ formas de elegir muestras sin piezas defectuosas y $\binom{100-k}{9} \binom{k}{1}$ formas de elegir muestras con exactamente una pieza defectuosa. En consecuencia la probabilidad $p(k)$ de que la partida pase la inspección es

$$p(k) = \binom{100-k}{10} \binom{k}{0} \binom{100}{10}^{-1} + \binom{100-k}{9} \binom{k}{1} \binom{100}{10}^{-1}.$$

Una cuenta sencilla muestra que para todo $k = 1, \dots, 90$ el cociente $\frac{p(k)}{p(k-1)}$ es menor que 1. Esto significa que a medida que aumenta la cantidad de piezas defectuosas en la partida, la probabilidad de aceptarla disminuye.

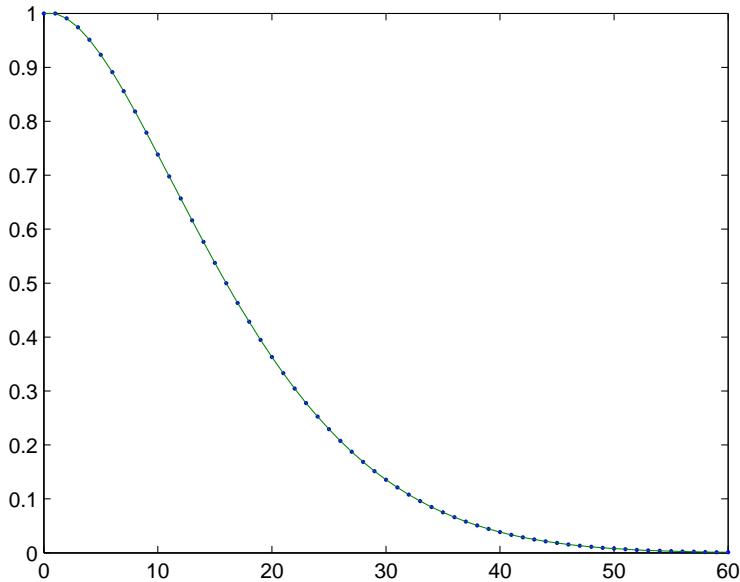


Figura 1: Gráfico de función $p(k)$.

¿Cuál es la máxima probabilidad de aceptar una partida de 100 que contenga más de 20 piezas defectuosas? Debido a que la función $p(k)$ es decreciente, dicha probabilidad es $p(20) \approx 0.3630$. \square

Ejemplo 3.14. Una planta de ensamblaje recibe un lote de $n = 100$ piezas de precisión, de las cuales una cantidad desconocida n_1 son defectuosas. Para controlar el lote se elige una muestra (sin reposición) de $r = 10$ piezas. Examinadas estas, resultan $k = 2$ defectuosas. ¿Qué se puede decir sobre la cantidad de piezas defectuosas en el lote?

Sabemos que de 10 piezas examinadas 2 son defectuosas y 8 no lo son. Por lo tanto, $2 \leq n_1 \leq 92$. Esto es todo lo que podemos decir con absoluta certeza. Podría suponerse que el lote contiene 92 piezas defectuosas. Partiendo de esa hipótesis, llegamos a la conclusión de que ha ocurrido un evento de probabilidad

$$\binom{8}{8} \binom{92}{2} \binom{100}{10}^{-1} = O(10^{-10}).$$

En el otro extremo, podría suponerse que el lote contiene exactamente 2 piezas defectuosas, en ese caso llegamos a la conclusión de que ha ocurrido un evento de probabilidad

$$\binom{98}{8} \binom{2}{2} \binom{100}{10}^{-1} = \frac{1}{110}.$$

Las consideraciones anteriores conducen a buscar el valor de n_1 que maximice la probabilidad

$$p(n_1) := \binom{100 - n_1}{8} \binom{n_1}{2} \binom{100}{10}^{-1},$$

puesto que para ese valor de n_1 nuestra observación tendría la mayor probabilidad de ocurrir. Para encontrar ese valor consideramos el cociente $\frac{p(n_1)}{p(n_1-1)}$. Simplificando los factoriales, obtenemos

$$\begin{aligned} \frac{p(n_1)}{p(n_1-1)} &= \frac{n_1(93-n_1)}{(n_1-2)(101-n_1)} > 1 \\ \iff n_1(93-n_1) &> (n_1-2)(101-n_1) \\ \iff n_1 < 20.2 &\iff n_1 \leq 20. \end{aligned}$$

Esto significa que cuando n_1 crece la sucesión $p(n_1)$ primero crece y después decrece; alcanza su máximo cuando $n_1 = 20$. Suponiendo que $n_1 = 20$, la probabilidad de que en una muestra de 10 piezas extraídas de un lote de 100 se observen 2 defectuosas es:

$$p(20) = \binom{80}{8} \binom{20}{2} \binom{100}{10}^{-1} \approx 0.318.$$

Aunque el verdadero valor de n_1 puede ser mayor o menor que 20, si se supone que $n_1 = 20$ se obtiene un resultado consistente con el sentido común que indicaría que los eventos observables deben tener “alta probabilidad”.

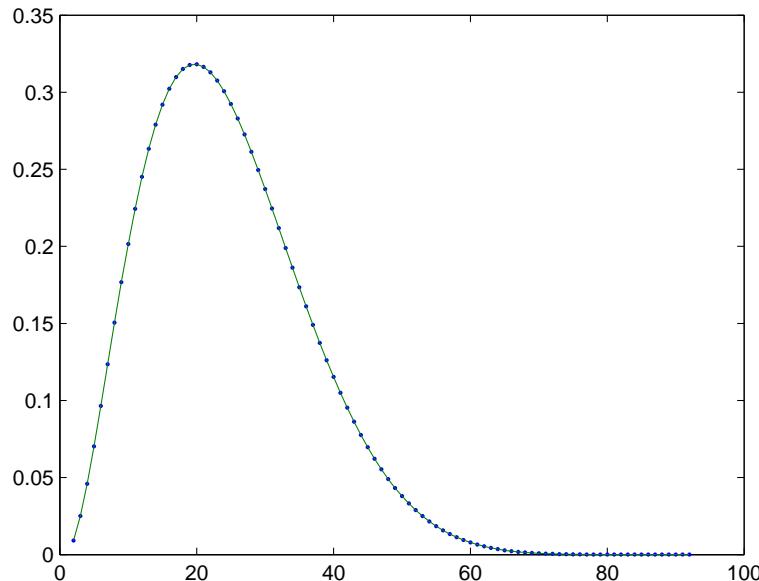


Figura 2: Gráfico de función $p(n_1)$. Observar que $\arg \max\{p(n_1) : 2 \leq n_1 \leq 92\} = 20$. \square

3.5.2. Estimación por captura y recaptura.

Para estimar la cantidad n de peces en un lago se puede realizar el siguiente procedimiento. En el primer paso se capturan n_1 peces, que luego de marcarlos se los deja en libertad. En el segundo paso se capturan r peces y se determina la cantidad k de peces marcados. La fórmula (21) permite hacer inferencias sobre la cantidad desconocida n .

Ejemplo 3.15 (Experimentos de captura y recaptura). Se capturan 1000 peces en un lago, se marcan con manchas rojas y se los deja en libertad. Después de un tiempo se hace una nueva captura de 1000 peces, y se encuentra que 100 tienen manchas rojas. ¿Qué conclusiones pueden hacerse sobre la cantidad de peces en el lago?

Suponemos que las dos capturas pueden considerarse como muestras aleatorias de la población total de peces en el lago. También vamos a suponer que la cantidad de peces en el lago no cambió entre las dos capturas.

Generalizamos el problema admitiendo tamaños muestrales arbitrarios. Sean

- n = el número (desconocido) de peces en el lago.
- n_1 = el número de peces en la primera captura. Estos peces juegan el rol de las bolas rojas.
- r = el número de peces en la segunda captura.
- k = el número de peces rojos en la segunda captura.
- $p_k(n)$ = la probabilidad de que la segunda captura contenga exactamente k peces rojos.

Con este planteo la probabilidad $p_k(n)$ se obtiene poniendo $n_2 = n - n_1$ en la fórmula (21):

$$p_k(n) = \binom{n_1}{k} \binom{n - n_1}{r - k} \binom{n}{r}^{-1}. \quad (22)$$

En la práctica n_1, r , y k pueden observarse, pero n es desconocido.

Notar que n es un número fijo que no depende del azar. Resultaría insensato preguntar por la probabilidad que n sea mayor que, digamos, 6000.

Sabemos que fueron capturados $n_1 + r - k$ peces diferentes, y por lo tanto $n \geq n_1 + r - k$. Esto es todo lo que podemos decir con absoluta certeza. En nuestro ejemplo tenemos $n_1 = r = 1000$ y $k = 100$, y podría suponerse que el lago contiene solamente 1900 peces. Sin embargo, partiendo de esa hipótesis, llegamos a la conclusión de que ha ocurrido un evento de probabilidad fantásticamente pequeña. En efecto, si se supone que hay un total de 1900 peces, la fórmula (22) muestra que la probabilidad de que las dos muestras de tamaño 1000 agoten toda la población es ,

$$\binom{1000}{100} \binom{900}{900} \binom{1900}{1000}^{-1} = \frac{(1000!)^2}{100! 1900!}$$

La fórmula de Stirling muestra que esta probabilidad es del orden de magnitud de 10^{-430} , y en esta situación el sentido común indica rechazar la hipótesis como irrazonable. Un razonamiento similar nos induce a rechazar la hipótesis de que n es muy grande, digamos, un millón.

Las consideraciones anteriores nos conducen a buscar el valor de n que maximice la probabilidad $p_k(n)$, puesto que para ese n nuestra observación tendría la mayor probabilidad de ocurrir. Para cualquier conjunto de observaciones n_1, r, k , el valor de n que maximiza la probabilidad $p_k(n)$ se denota por \hat{n}_{mv} y se llama el *estimador de máxima verosimilitud* de n . Para

encontrar \hat{n}_{mv} consideramos la proporción

$$\begin{aligned} \frac{p_k(n)}{p_k(n-1)} &= \frac{(n-n_1)(n-r)}{(n-n_1-r+k)n} > 1 \\ \iff &(n-n_1)(n-r) > (n-n_1-r+k)n \\ \iff &n^2 - nn_1 - nr + n_1r > n^2 - nn_1 - nr + nk \\ \iff &n < \frac{n_1r}{k}. \end{aligned}$$

Esto significa que cuando n crece la sucesión $p_k(n)$ primero crece y después decrece; alcanza su máximo cuando n es el mayor entero menor que $\frac{n_1r}{k}$, así que \hat{n}_{mv} es aproximadamente igual a $\frac{n_1r}{k}$. En nuestro ejemplo particular el estimador de máxima verosimilitud del número de peces en el lago es $\hat{n}_{mv} = 10000$.

El verdadero valor de n puede ser mayor o menor, y podemos preguntar por los límites entre los que resulta razonable esperar que se encuentre n . Para esto testeamos la hipótesis que n sea menor que 8500. Sustituimos en (22) $n = 8500$, $n_1 = r = 1000$, y calculamos la probabilidad que la segunda muestra contenga 100 o menos peces rojos. Esta probabilidad es $p = p_0 + p_1 + \dots + p_{100}$. Usando una computadora encontramos que $p \approx 0.04$. Similarmente, si $n = 12.000$, la probabilidad que la segunda muestra contenga 100 o más peces rojos esta cerca de 0.03. Esos resultados justificarían la apuesta de que el verdadero número n de peces se encuentra en algún lugar entre 8500 y 12.000.

□

Ejercicios adicionales

- 10.** Un estudiante de ecología va a una laguna y captura 60 escarabajos de agua, marca cada uno con un punto de pintura y los deja en libertad. A los pocos días vuelve y captura otra muestra de 50, encontrando 12 escarabajos marcados. ¿Cuál sería su mejor apuesta sobre el tamaño de la población de escarabajos de agua en la laguna?

4. Mecánica Estadística

El espacio se divide en una gran cantidad, n , de pequeñas regiones llamadas celdas. Se considera un sistema mecánico compuesto por r partículas que se distribuyen al azar entre las n celdas. ¿Cuál es la distribución de las partículas en las celdas? La respuesta depende de lo que se considere un evento elemental.

1. *Estadística de Maxwell-Boltzmann.* Suponemos que todas las partículas son distintas y que todas las ubicaciones de las partículas son igualmente posibles. Un evento elemental está determinado por la r -upla (x_1, x_2, \dots, x_r) , donde x_i es el número de la celda en la que cayó la partícula i . Puesto que cada x_i puede tomar n valores distintos, el número de tales r -uplas es n^r . La probabilidad de un evento elemental es $1/n^r$.
2. *Estadística de Bose-Einstein.* Las partículas son indistinguibles. De nuevo, todas las ubicaciones son igualmente posibles. Un evento elemental está determinado por la n -upla

(r_1, \dots, r_n) , donde $r_1 + \dots + r_n = r$ y r_i es la cantidad de partículas en la i -ésima celda, $1 \leq i \leq n$. La cantidad de tales n -uplas se puede calcular del siguiente modo: a cada n - upla (r_1, r_2, \dots, r_n) la identificamos con una sucesión de unos y ceros s_1, \dots, s_{r+n-1} con unos en las posiciones numeradas $r_1 + 1, r_1 + r_2 + 2, \dots, r_1 + r_2 + \dots + r_{n-1} + n - 1$ (hay $n - 1$ de ellas) y ceros en las restantes posiciones. La cantidad de tales sucesiones es igual al número de combinaciones de $r + n - 1$ cosas tomadas de a $n - 1$ por vez. La probabilidad de un evento elemental es $1/{r+n-1 \choose n-1}$.

3. *Estadística de Fermi-Dirac.* En este caso $r < n$ y cada celda contiene a lo sumo una partícula. La cantidad de eventos elementales es ${n \choose r}$. La probabilidad de un evento elemental es $1/{n \choose r}$.

Ejemplo 4.1. Se distribuyen 5 partículas en 10 celdas numeradas 1, 2, ..., 10. Calcular, para cada una de las tres estadísticas, la probabilidad de que las celdas 8, 9 y 10 no tengan partículas y que las celdas 6 y 7 tengan exactamente una partícula cada una.

1. *Maxwell-Boltzmann.* Las bolas son distinguibles y todas las configuraciones diferentes son equiprobables. La probabilidad de cada configuración $(x_1, \dots, x_5) \in \{1, \dots, 10\}^5$, donde x_i indica la celda en que se encuentra la partícula i , es $1/10^5$.

¿De qué forma podemos obtener las configuraciones deseadas? Primero elegimos (en orden) las 2 bolas que van a ocupar la celdas 6 y 7 (hay 5×4 formas diferentes de hacerlo) y luego elegimos entre las celdas 1, 2, 3, 4, 5 las ubicaciones de las 3 bolas restantes (hay 5^3 formas diferentes de hacerlo). Por lo tanto, su cantidad es $5 \times 4 \times 5^3$ y la probabilidad de observarlas es

$$p = \frac{5 \times 4 \times 5^3}{10^5} = \frac{1}{5 \times 2^3} = \frac{1}{40} = 0.025.$$

2. *Bose-Einstein.* Las partículas son indistinguibles y todas las configuraciones distintas son equiprobables. La probabilidad de cada configuración (r_1, \dots, r_{10}) , donde $r_1 + \dots + r_{10} = 5$ y r_i es la cantidad de partículas en la i -ésima celda, es $1/{14 \choose 9}$.

Las configuraciones deseadas son de la forma $(r_1, \dots, r_5, 1, 1, 0, 0, 0)$, donde $r_1 + \dots + r_5 = 3$, su cantidad es igual a la cantidad de configuraciones distintas que pueden formarse usando 3 ceros y 4 unos. Por lo tanto, su cantidad es ${7 \choose 3}$ y la probabilidad de observarlas es

$$p = {7 \choose 3} {14 \choose 9}^{-1} = \frac{35}{2002} \approx 0.0174....$$

3. *Fermi-Dirac.* Las partículas son indistinguibles, ninguna celda puede contener más de una partícula y todas las configuraciones distintas son equiprobables. La probabilidad de cada configuración es $1/{10 \choose 5}$.

Las configuraciones deseadas se obtienen eligiendo tres de las cinco celdas 1, 2, 3, 4, 5 para ubicar las tres partículas que no están en las celdas 6 y 7. Por lo tanto, su cantidad es ${5 \choose 3}$ y la probabilidad de observarlas es

$${5 \choose 3} {10 \choose 5}^{-1} = \frac{10}{252} \approx 0.0396....$$

□

Ejemplo 4.2. Calcular para cada una de las tres estadísticas mencionadas, la probabilidad de que una celda determinada (p.ej., la número 1) no contenga partícula.

En cada uno de los tres casos la cantidad de eventos elementales favorables es igual a la cantidad de ubicaciones de las partículas en $n - 1$ celdas. Por lo tanto, designando por p_{MB}, p_{BE}, p_{FD} las probabilidades del evento especificado para cada una de las estadísticas (siguiendo el orden de exposición), tenemos que

$$\begin{aligned} p_{MB} &= \frac{(n-1)^r}{n^r} = \left(1 - \frac{1}{n}\right)^r, \\ p_{BE} &= \binom{r+n-2}{n-2} \binom{r+n-1}{n-1}^{-1} = \frac{n-1}{N+n-1}, \\ p_{FD} &= \binom{n-1}{r} \binom{n}{r}^{-1} = 1 - \frac{r}{n}. \end{aligned}$$

Si $r/n = \lambda$ y $n \rightarrow \infty$, entonces

$$p_{MB} = e^{-\lambda}, \quad p_{BE} = \frac{1}{1+\lambda}, \quad p_{FD} = 1 - \lambda.$$

Si λ es pequeño, esas probabilidades coinciden hasta $O(\lambda^2)$. El número λ caracteriza la “densidad promedio” de las partículas. □

Ejercicios adicionales

11. Utilizando la estadística de Maxwell-Boltzmann construir un mecanismo aleatorio para estimar el número e .

4.1. Algunas distribuciones relacionadas con la estadística de Maxwell-Boltzmann

Se distribuyen r partículas en n celdas y cada una de las n^r configuraciones tiene probabilidad n^{-r} .

4.1.1. Cantidad de partículas por celda: la distribución binomial

Cantidad de partículas en una celda específica. Para calcular la probabilidad, $p_{MB}(k)$, de que una celda específica contenga exactamente k partículas ($k = 0, 1, \dots, r$) notamos que las k partículas pueden elegirse de $\binom{r}{k}$ formas, y las restantes $r - k$ partículas pueden ubicarse en las restantes $n - 1$ celdas de $(n - 1)^{r-k}$ formas. Resulta que

$$p_{MB}(k) = \binom{r}{k} (n-1)^{r-k} \frac{1}{n^r}$$

Dicho en palabras, en la estadística de Maxwell-Boltzmann la probabilidad de que una celda dada contenga exactamente k partículas está dada por la distribución Binomial $(r, \frac{1}{n})$ definida por

$$p(k) := \binom{r}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{r-k}, \quad 0 \leq k \leq r. \quad (23)$$

□

Cantidad de partículas más probable en una celda específica. La cantidad más probable de partículas en una celda específica es el entero ν tal que

$$\frac{(r-n+1)}{n} < \nu \leq \frac{(r+1)}{n}. \quad (24)$$

Para ser más precisos:

$$p_{MB}(0) < p_{MB}(1) < \dots < p_{MB}(\nu-1) \leq p_{MB}(\nu) > p_{MB}(\nu+1) > \dots > p_{MB}(r).$$

Demostración. (Ejercicio.)

□

4.1.2. Forma límite: la distribución de Poisson

Forma límite. Si $n \rightarrow \infty$ y $r \rightarrow \infty$ de modo que la cantidad promedio $\lambda = r/n$ de partículas por celda se mantiene constante, entonces

$$p_{MB}(k) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

Dicho en palabras, la *forma límite* de la estadística de Maxwell-Boltzmann es la *distribución de Poisson de media* λ definida por

$$p(k) := e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots \quad (25)$$

Demostración. Primero observamos que:

$$\begin{aligned} \binom{r}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{r-k} &= \frac{r!}{k!(r-k)!} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{r-k} \\ &= \frac{1}{k!} \left(\frac{1}{n}\right)^k \left(\frac{n-1}{n}\right)^{-k} \frac{r!}{(r-k)!} \left(1 - \frac{1}{n}\right)^r \\ &= \frac{1}{k!} \frac{1}{(n-1)^k} \frac{r!}{(r-k)!} \left(1 - \frac{1}{n}\right)^r. \end{aligned} \quad (26)$$

Reemplazando en (26) $r = \lambda n$ obtenemos:

$$\begin{aligned} \binom{\lambda n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{\lambda n - k} &= \frac{1}{k!} \frac{1}{(n-1)^k} \frac{(\lambda n)!}{(\lambda n - k)!} \left(1 - \frac{1}{n}\right)^{\lambda n} \\ &= \left[\left(1 - \frac{1}{n}\right)^n \right]^\lambda \frac{1}{k!} \frac{1}{(n-1)^k} \frac{(\lambda n)!}{(\lambda n - k)!} \\ &\sim e^{-\lambda} \frac{1}{k!} \left(\frac{1}{(n-1)^k} \frac{(\lambda n)!}{(\lambda n - k)!} \right). \end{aligned} \quad (27)$$

Para estimar el último factor del lado derecho de (27) utilizamos la fórmula de Stirling $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$:

$$\begin{aligned}
\frac{1}{(n-1)^k} \frac{(\lambda n)!}{(\lambda n - k)!} &\sim \frac{1}{(n-1)^k} \frac{\sqrt{2\pi} (\lambda n)^{\lambda n + \frac{1}{2}} e^{-\lambda n}}{\sqrt{2\pi} (\lambda n - k)^{(\lambda n - k) + \frac{1}{2}} e^{-(\lambda n - k)}} \\
&= \frac{1}{(n-1)^k} \frac{(\lambda n)^{\lambda n + \frac{1}{2}} e^{-k}}{(\lambda n - k)^{(\lambda n - k) + \frac{1}{2}}} \\
&= \left(\frac{\lambda n - k}{n-1} \right)^k \left(\frac{\lambda n}{\lambda n - k} \right)^{\lambda n + \frac{1}{2}} e^{-k} \\
&\sim \lambda^k e^{-k} \left[\left(1 - \frac{k}{\lambda n} \right)^{\lambda n + \frac{1}{2}} \right]^{-1} \\
&\sim \lambda^k.
\end{aligned} \tag{28}$$

De (26), (27) y (28) resulta que

$$\binom{r}{k} \left(\frac{1}{n} \right)^k \left(1 - \frac{1}{n} \right)^{r-k} \sim e^{-\lambda} \frac{\lambda^k}{k!}. \tag{29}$$

□

4.2. Algunas distribuciones relacionadas con la estadística de Bose-Einstein

Se distribuyen r partículas indistinguibles en n celdas y cada una de las $\binom{r+n-1}{n-1}$ configuraciones tiene probabilidad $1/\binom{r+n-1}{n-1}$.

4.2.1. Cantidad de partículas por celda

Cantidad de partículas en una celda específica. Para calcular la probabilidad, $p_{BE}(k)$, de que una celda específica contenga exactamente k partículas ($k = 0, 1, \dots, r$) fijamos k de los r ceros y 1 de los $n-1$ unos para representar que hay k partículas en la urna específica. La cantidad de configuraciones distintas que pueden formarse con los restantes $r-k$ ceros y $n-2$ unos es $\binom{r-k+n-2}{n-2}$. Resulta que

$$p_{BE}(k) = \binom{r-k+n-2}{n-2} \binom{r+n-1}{n-1}^{-1}. \tag{30}$$

Cantidad de partículas más probable en una celda específica. Cuando $n > 2$ la cantidad más probable de partículas en una celda específica es 0 o más precisamente $p_{BE}(0) > p_{BE}(1) > \dots$

Demostración. (Ejercicio.)

□

4.2.2. Forma límite: la distribución de Geométrica

Forma límite. Si $n \rightarrow \infty$ y $r \rightarrow \infty$ de modo que la cantidad promedio $\lambda = r/n$ de partículas por celda se mantiene constante, entonces

$$p_{BE}(k) \rightarrow \frac{\lambda^k}{(1 + \lambda)^{k+1}}.$$

Dicho en palabras, la *forma límite* de la estadística de Bose-Einstein es la *distribución geométrica de parámetro* $\frac{1}{1+\lambda}$ definida por

$$p(k) := \left(1 - \frac{1}{1 + \lambda}\right)^k \frac{1}{1 + \lambda}, \quad k = 0, 1, 2, \dots$$

Demostración. Primero observamos que:

$$\begin{aligned} \binom{r-k+n-2}{n-2} \binom{r+n-1}{n-1}^{-1} &= \frac{(r-k+n-2)!}{(n-2)!(r-k)!} \frac{(n-1)!r!}{(r+n-1)!} \\ &= \frac{(n-1)!}{(n-2)!} \frac{r!}{(r-k)!} \frac{(r-k+n-2)!}{(r+n-1)!}. \end{aligned} \quad (31)$$

Reemplazando en el lado derecho de (31) $r = \lambda n$ obtenemos:

$$\frac{(n-1)!}{(n-2)!} \frac{(\lambda n)!}{(\lambda n - k)!} \frac{(\lambda n - k + n - 2)!}{(\lambda n + n - 1)!} \quad (32)$$

Para estimar los factores que intervienen en (32) utilizamos la fórmula de Stirling $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$:

$$\begin{aligned} \frac{(n-1)^{n-1+\frac{1}{2}} e^{-n+1}}{(n-2)^{n-2+\frac{1}{2}} e^{-n+2}} &\sim (n-2)e^{-1} \left[\left(1 - \frac{1}{n-1}\right)^{n-1} \right]^{-1} \\ &\sim n-2 \sim n, \end{aligned} \quad (33)$$

$$\begin{aligned} \frac{(\lambda n)^{\lambda n+\frac{1}{2}} e^{-\lambda n}}{(\lambda n - k)^{\lambda n-k+\frac{1}{2}} e^{-\lambda n+k}} &\sim (\lambda n - k)^k e^{-k} \left[\left(1 - \frac{k}{\lambda n}\right)^{\lambda n} \right]^{-1} \\ &\sim (\lambda n - k)^k \sim \lambda^k n^k, \end{aligned} \quad (34)$$

$$\begin{aligned} \frac{(\lambda n - k + n - 2)^{\lambda n-k+n-2+\frac{1}{2}} e^{-\lambda n+k-n+2}}{(\lambda n + n - 1)^{\lambda n+n-1+\frac{1}{2}} e^{-\lambda n-n+1}} &\sim (\lambda n - k + n - 2)^{-k-1} e^{k+1} \\ &\quad \times \left(1 - \frac{k+1}{\lambda n + n - 1}\right)^{\lambda n+n-1} \\ &\sim (\lambda n - k + n - 2)^{-k-1} \\ &\sim \frac{1}{(1 + \lambda)^{k+1} n^{k+1}}. \end{aligned} \quad (35)$$

De (31), (32), (33), (34) y (35) resulta que

$$\binom{r-k+n-2}{n-2} \binom{r+n-1}{n-1}^{-1} \sim \frac{\lambda^k}{(1 + \lambda)^k}. \quad (36)$$

□

Ejercicios adicionales

12. Considerando la estadística de Maxwell-Boltzmann para la distribución aleatoria de r partículas en n celdas demostrar que la cantidad de partículas más probable en una celda determinada es la parte entera de $\frac{r+1}{n}$.

13. Considerando la estadística de Bose-Einstein para la distribución aleatoria de r partículas (indistinguibles) en $n > 2$ celdas demostrar que la cantidad de partículas más probable en una celda determinada es 0.

4.3. Tiempos de espera

Consideramos una vez más el experimento conceptual de ubicar aleatoriamente partículas (distinguibles) en n celdas. Solo que ahora no fijamos la cantidad r de partículas y ubicamos las partículas una por una hasta que ocurra alguna situación prescrita. Analizaremos dos situaciones:

- (i) Ubicar partículas hasta que alguna se ubique en una celda ocupada previamente.
- (ii) Fijada una celda, ubicar partículas hasta que alguna ocupe la celda.

Situación (i). Usamos símbolos de la forma (j_1, j_2, \dots, j_r) para indicar que la primera, la segunda,... y la r -ésima partícula están ubicadas en las celdas j_1, j_2, \dots, j_r y que el proceso culmina en el paso r . Esto significa que las j_i son enteros entre 1 y n ; que las j_1, j_2, \dots, j_{r-1} son todas diferentes y que j_r es igual a una de ellas. Toda configuración de ese tipo representa un punto muestral. Los posibles valores de r son $2, 3, \dots, n + 1$.

Para un r fijo el conjunto de todos los puntos muestrales (j_1, j_2, \dots, j_r) representa el evento que el proceso termina en el r -ésimo paso. Los números j_1, j_2, \dots, j_{r-1} pueden elegirse de $(n)_{r-1}$ formas diferentes; j_r podemos elegir uno de los $r - 1$ números j_1, j_2, \dots, j_{r-1} . Por lo tanto la probabilidad de que el proceso termine en el r -ésimo paso es

$$p_r = \frac{(n)_{r-1}(r-1)}{n^r}. \quad (37)$$

□

Situación (ii). Usamos símbolos de la forma (j_1, j_2, \dots, j_r) para indicar que la primera, la segunda,... y la r -ésima partícula están ubicadas en las celdas j_1, j_2, \dots, j_r y que el proceso culmina en el paso r . Las r -uplas (j_1, j_2, \dots, j_r) están sujetas a la condición de que los números j_1, j_2, \dots, j_{r-1} son diferentes de un número prescrito $a \leq n$, y $j_r = a$.

Para un r fijo el conjunto de todos los puntos muestrales (j_1, j_2, \dots, j_r) representa el evento que el proceso termina en el r -ésimo paso. Los números j_1, j_2, \dots, j_{r-1} pueden elegirse de $(n-1)^{r-1}$ formas diferentes; j_r debe ser a . Por lo tanto la probabilidad de que el proceso termine en el r -ésimo paso es

$$p_r = \frac{(n-1)^{r-1}}{n^r}. \quad (38)$$

□

5. Bibliografía consultada

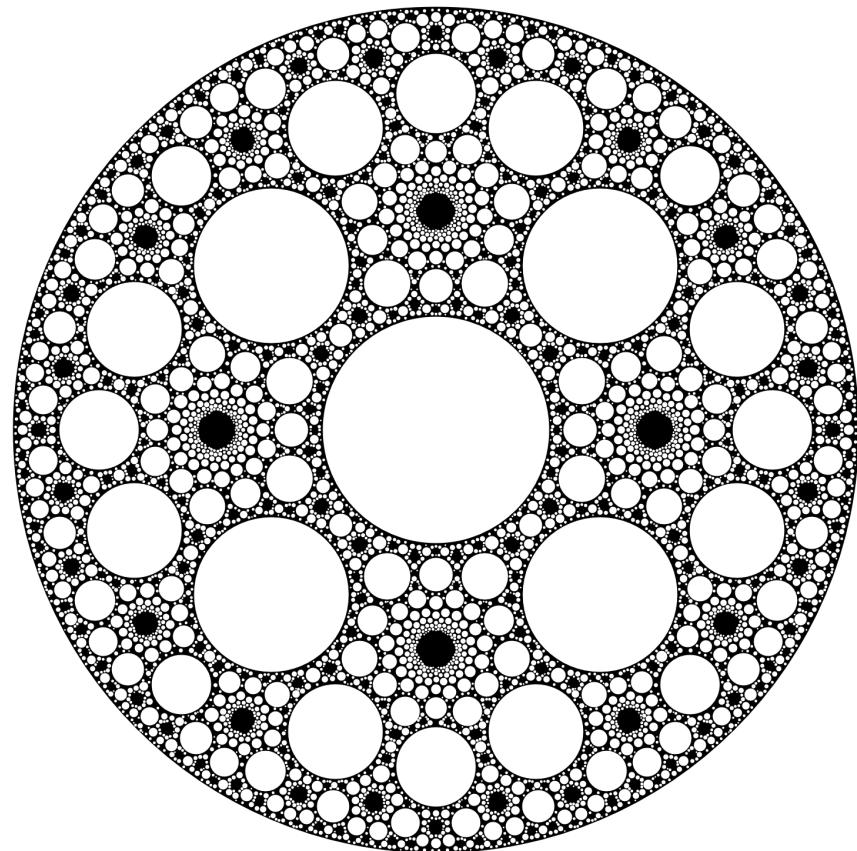
Para redactar estas notas se consultaron los siguientes libros:

1. Bertsekas, D. P., Tsitsiklis, J. N.: Introduction to Probability. M.I.T. Lecture Notes. (2000)
2. Brémaud, P.: An Introduction to Probabilistic Modeling. Springer, New York. (1997)
3. Durrett, R. Elementary Probability for Applications. Cambridge University Press, New York. (2009)
4. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons, New York. (1957)
5. Ferrari, P.: Passeios aleatórios e redes eletricas. Instituto de Matemática Pura e Aplicada. Rio de Janeiro. (1987)
6. Grinstead, C. M. & Snell, J. L. Introduction to Probability. American Mathematical Society. (1997)
7. Kolmogorov, A. N.: Foundations of the Theory of Probability. Chelsea Publishing Co., New York. (1956)
8. Kolmogorov, A. N.: The Theory of Probability. Mathematics. Its Content, Methods, and Meaning. Vol 2. The M.I.T. Press, Massachusetts. (1963) pp. 229-264.
9. Meester, R.: A Natural Introduction to Probability Theory. Birkhauser, Berlin. (2008)
10. Meyer, P. L.: Introductory Probability and Statistical Applications. Addison-Wesley, Massachusetts. (1972)
11. Ross, S. M: Introduction to Probability and Statistics for Engineers and Scientists. Elsevier Academic Press, San Diego. (2004)
12. Skorokhod, A. V.: Basic Principles and Applications of Probability Theory. Springer-Verlag, Berlin. (2005)
13. Soong, T. T.: Fundamentals of Probability and Statistics for Engineers. John Wiley & Sons Ltd. (2004)
14. Stoyanov, J.: Counterexamples in Probability. John Wiley & Sons. (1997)

Probabilidad Condicional, Independencia Estocástica
Algunos modelos probabilísticos
(Borradores, Curso 23)

Sebastian Grynberg

18-20 de marzo 2013



*“No importa lo que yo piense.
Es lo que tú piensas lo que es relevante.”*

(Dr. House)

Índice

1. Probabilidad Condicional	3
1.1. Probabilidad Condicional	3
1.2. Fórmula de probabilidad total	4
1.3. Regla de Bayes	7
2. Independencia estocástica	10
3. Modelos discretos	11
4. Modelos continuos	14
4.1. Puntos al azar sobre un segmento. La distribución uniforme	14
4.2. Geometría y probabilidad	15
4.3. Paradoja de Bertrand	17
4.4. De las masas puntuales a la masa continua	18
5. Bibliografía consultada	20

1. Probabilidad Condicional

1.1. Probabilidad Condicional

Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad.

Definición 1.1 (Probabilidad condicional). Sea $A \subset \Omega$ un evento de probabilidad positiva. Para cada evento B definimos

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}. \quad (1)$$

La cantidad definida en (1) se llama la *probabilidad condicional de B dado que ocurrió A* .

Nota Bene (La probabilidad condicional induce una medida de probabilidad sobre los eventos aleatorios). Valen las siguientes propiedades:

1. Para cada $B \in \mathcal{A}$, $\mathbb{P}(B|A) \geq 0$;
2. $\mathbb{P}(\Omega|A) = 1$;
3. Si los eventos B y C no tienen elementos en común, entonces

$$\mathbb{P}(B \cup C|A) = \mathbb{P}(B|A) + \mathbb{P}(C|A).$$

4. Para cada sucesión decreciente de eventos $B_1 \supset B_2 \supset \dots$ tal que $\bigcap_{n=1}^{\infty} B_n = \emptyset$ vale que $\lim_{n \rightarrow \infty} \mathbb{P}(B_n|A) = 0$.

Comparando las propiedades 1-4 con los axiomas I-IV, se concluye que la función $\mathbb{P}(\cdot|A) : \mathcal{A} \rightarrow \mathbb{R}$ es una medida de probabilidad sobre los eventos aleatorios. Por lo tanto, todos los resultados generales referidos a la propiedades de $\mathbb{P}(\cdot)$ también valen para la probabilidad condicional $\mathbb{P}(\cdot|A)$. \square

Ejemplo 1.2. Se lanza un dado equilibrado. Sabiendo que el resultado del dado no superó al 4, cuál es la probabilidad condicional de haber obtenido un 3? Denotando mediante A al evento “el resultado no supera al 4” y mediante B el evento “el resultado es 3”. Tenemos que $\mathbb{P}(A) = 4/6$, $\mathbb{P}(B) = 1/6$ y $\mathbb{P}(A \cap B) = \mathbb{P}(A) = 1/6$. Así

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{1/6}{4/6} = \frac{1}{4},$$

lo que intuitivamente tiene sentido (¿por qué?). \square

Probabilidad compuesta. De la definición de la probabilidad condicional del evento B dado que ocurrió el evento A resulta inmediatamente la siguiente fórmula

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A). \quad (2)$$

denominada *regla del producto*.

El siguiente Teorema generaliza la regla del producto (2) y se obtiene por inducción.

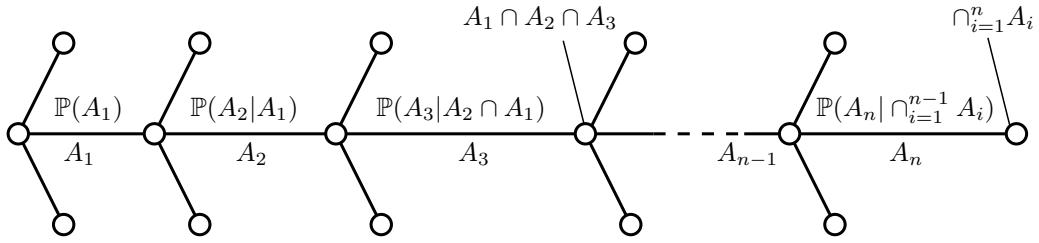


Figura 1: **Ilustración de la regla del producto.** El evento $\cap_{i=1}^n A_i$ tiene asociada una única trayectoria sobre un árbol que describe la historia de un experimento aleatorio realizado por etapas sucesivas. Las aristas de esta trayectoria corresponden a la ocurrencia sucesiva de los eventos A_1, A_2, \dots, A_n y sobre ellas registramos la correspondiente probabilidad condicional. El nodo final de la trayectoria corresponde al evento $\cap_{i=1}^n A_i$ y su probabilidad se obtiene multiplicando las probabilidades condicionales registradas a lo largo de las aristas de la trayectoria: $\mathbb{P}(\cap_{i=1}^n A_i) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_2 \cap A_1) \cdots \mathbb{P}(A_n|\cap_{i=1}^{n-1} A_i)$. Notar que cada nodo intermedio a lo largo de la trayectoria también corresponde a un evento intersección y su probabilidad se obtiene multiplicando las probabilidades condicionales registradas desde el inicio de la trayectoria hasta llegar al nodo. Por ejemplo, el evento $A_1 \cap A_2 \cap A_3$ corresponde al nodo indicado en la figura y su probabilidad es $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2)$.

Teorema 1.3 (Regla del producto). Suponiendo que todos los eventos condicionantes tienen probabilidad positiva, tenemos que

$$\mathbb{P}(\cap_{i=1}^n A_i) = \mathbb{P}(A_n | \cap_{i=1}^{n-1} A_i) \cdots \mathbb{P}(A_3 | A_1 \cap A_2) \mathbb{P}(A_2 | A_1) \mathbb{P}(A_1). \quad (3)$$

Ejemplo 1.4. Una urna contiene 5 bolas rojas y 10 bolas negras. Se extraen dos bolas al azar sin reposición. ¿Cuál es la probabilidad que ambas bolas sean negras?

Sean N_1 y N_2 los eventos definidos por “la primera bola extraída es negra” y “la segunda bola extraída es negra”, respectivamente. Claramente $\mathbb{P}(N_1) = 10/15$. Para calcular $\mathbb{P}(N_2|N_1)$ observamos que si ocurrió N_1 , entonces solo 9 de las 14 bolas restantes en la urna son negras. Así $\mathbb{P}(N_2|N_1) = 9/14$ y

$$\mathbb{P}(N_2 \cap N_1) = \mathbb{P}(N_2|N_1)\mathbb{P}(N_1) = \frac{10}{15} \cdot \frac{9}{14} = \frac{3}{7}.$$

□

1.2. Fórmula de probabilidad total

Teorema 1.5 (Fórmula de probabilidad total). Sea A_1, A_2, \dots una sucesión de eventos disjuntos dos a dos tal que $\bigcup_{n \geq 1} A_n = \Omega$. Para cada $B \in \mathcal{A}$ vale la siguiente fórmula

$$\mathbb{P}(B) = \sum_{n \geq 1} \mathbb{P}(B|A_n)\mathbb{P}(A_n), \quad (4)$$

denominada *fórmula de probabilidad total*¹.

¹Rigurosamente, $\mathbb{P}(B|A_n)$ está definida cuando $\mathbb{P}(A_n) > 0$, por lo cual en la fórmula (4) interpretaremos que $\mathbb{P}(B|A_n)\mathbb{P}(A_n) = 0$ cuando $\mathbb{P}(A_n) = 0$.

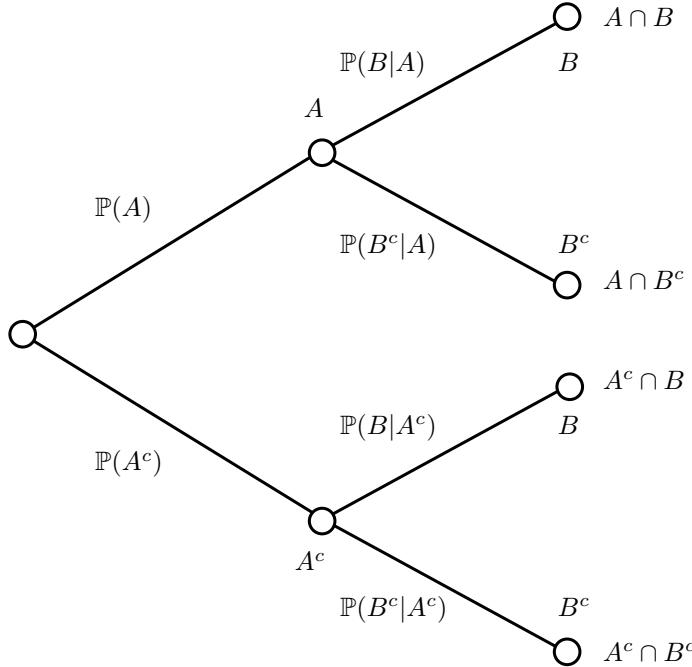


Figura 2: **Ilustración de la fórmula de probabilidad total.** Un experimento de dos etapas binarias y su correspondiente diagrama de árbol. La primera ramificación (de izquierda a derecha) se basa en el resultado de la primer etapa del experimento (A o A^c) y la segunda en su resultado final (B o B^c). Multiplicando las probabilidades registradas a lo largo de cada trayectoria se obtiene la probabilidad del evento intersección representado por el nodo final. Sumando las probabilidades de las trayectorias que corresponden al evento B se obtiene: $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)$.

Demostración de la fórmula de probabilidad total. De la identidad de conjuntos

$$B = B \cap \Omega = B \cap \left(\bigcup_{n \geq 1} A_n \right) = \bigcup_{n \geq 1} (B \cap A_n)$$

y la σ -aditividad de la medida de probabilidad \mathbb{P} se deduce que

$$\mathbb{P}(B) = \sum_{n=1}^{\infty} \mathbb{P}(B \cap A_n).$$

Si $\mathbb{P}(A_n) = 0$, $\mathbb{P}(B \cap A_n) = 0$ porque $B \cap A_n \subset A_n$. Si $\mathbb{P}(A_n) > 0$, entonces $\mathbb{P}(B \cap A_n) = \mathbb{P}(B|A_n)\mathbb{P}(A_n)$. \square

Nota Bene: Cálculo mediante condicionales. Si se dispone de una colección de eventos A_1, A_2, \dots de los cuales uno y solamente uno debe ocurrir, la fórmula de probabilidad total (4) permite calcular la probabilidad de cualquier evento B condicionando a saber cuál de los eventos A_i ocurrió. Más precisamente, la fórmula (4) establece que la probabilidad $\mathbb{P}(B)$ es igual al promedio ponderado de las probabilidades condicionales $\mathbb{P}(B|A_i)$ donde cada término

se pondera por la probabilidad del evento sobre el que se condicionó. Esta fórmula es útil debido a que a veces es más fácil evaluar las probabilidades condicionales $\mathbb{P}(B|A_i)$ que calcular directamente la probabilidad $\mathbb{P}(B)$. \square

Ejemplo 1.6 (Experimentos de dos etapas). La primera etapa del experimento produce una partición A_1, A_2, \dots del espacio muestral Ω . La segunda etapa produce el evento B . La fórmula (4) se utiliza para calcular la probabilidad de B .

Ejemplo 1.7. Una urna contiene 5 bolas rojas y 10 bolas negras. Se extraen dos bolas sin reposición. ¿Cuál es la probabilidad de que la segunda bola sea negra?

El espacio muestral de este experimento aleatorio se puede representar como las trayectorias a lo largo de un árbol como se muestra en la Figura 3.

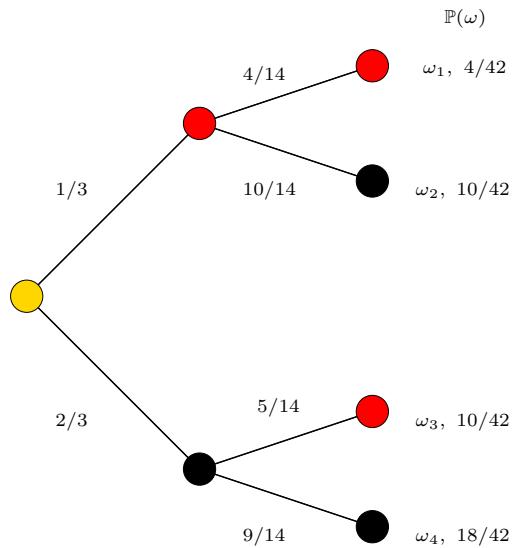


Figura 3: Observando el árbol se deduce que la probabilidad de que la segunda bola sea negra es: $\frac{1}{3} \cdot \frac{10}{14} + \frac{2}{3} \cdot \frac{9}{14} = \frac{2}{3}$.

Formalmente, el problema se resuelve mediante la fórmula de probabilidad total. Sean N_i y R_i los eventos definidos por “la i -ésima bola extraída es negra” y “la i -ésima bola extraída es roja”, respectivamente ($i = 1, 2$). Vale que

$$\mathbb{P}(N_1) = \frac{10}{15}, \quad \mathbb{P}(R_1) = \frac{5}{15}, \quad \mathbb{P}(N_2|R_1) = \frac{10}{14}, \quad \mathbb{P}(N_2|N_1) = \frac{9}{14}.$$

Usando la fórmula de probabilidad total obtenemos

$$\begin{aligned} \mathbb{P}(N_2) &= \mathbb{P}(N_2 \cap R_1) + \mathbb{P}(N_2 \cap N_1) \\ &= \mathbb{P}(N_2|R_1)\mathbb{P}(R_1) + \mathbb{P}(N_2|N_1)\mathbb{P}(N_1) \\ &= \frac{10}{14} \cdot \frac{1}{3} + \frac{9}{14} \cdot \frac{2}{3} = \frac{28}{42} = \frac{2}{3}. \end{aligned}$$

\square

1.3. Regla de Bayes

Primera versión de la regla de Bayes. Sean A y B dos eventos de probabilidad positiva. De la regla del producto (2) y su análoga $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$ se obtiene la siguiente fórmula importante

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}, \quad (5)$$

que contiene lo esencial del Teorema de Bayes.

Ejemplo 1.8. Un test de sangre es 95 % efectivo para detectar una enfermedad cuando una persona realmente la padece. Sin embargo, el test también produce un “falso positivo” en el 1 % de las personas saludables testeadas. Si el 0,5 % de la población padece la enfermedad, cuál es la probabilidad de que una persona tenga la enfermedad si su test resultó positivo?

Sea A el evento definido por “*la persona testeada tiene la enfermedad*” y sea B el evento definido por “*el resultado de su test es positivo*”. La probabilidad que nos interesa es $\mathbb{P}(A|B)$ y se puede calcular de la siguiente manera. Sabemos que

$$\mathbb{P}(A) = 0.005, \quad \mathbb{P}(A^c) = 0.995,$$

$$\mathbb{P}(B|A) = 0.95, \quad \mathbb{P}(B|A^c) = 0.01,$$

y usando esa información queremos calcular

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

El numerador, $\mathbb{P}(A \cap B)$, se puede calcular mediante la regla del producto

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) = (0.95)(0.005)$$

y el denominador, $\mathbb{P}(B)$, se puede calcular usando la fórmula de probabilidad total

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c) = (0.95)(0.005) + (0.01)(0.995).$$

Por lo tanto,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)} = \frac{95}{294} \approx 0.323.$$

En otras palabras, sólo el 32 % de aquellas personas cuyo test resultó positivo realmente tienen la enfermedad. \square

Teorema 1.9 (Bayes). Sean A_1, A_2, \dots , eventos disjuntos dos a dos y tales que $\bigcup_{n \geq 1} A_n = \Omega$.

Sea B un evento de probabilidad positiva. Entonces,

$$\mathbb{P}(A_n|B) = \frac{\mathbb{P}(B|A_n)\mathbb{P}(A_n)}{\sum_{k \geq 1} \mathbb{P}(B|A_k)\mathbb{P}(A_k)}, \quad n \geq 1. \quad (6)$$

Si los eventos A_1, A_2, \dots se llaman “hipótesis”, la fórmula (6) se considera como la probabilidad de ocurrencia de la hipótesis A_n sabiendo que ocurrió el evento B . En tal caso, $\mathbb{P}(A_n)$ es la probabilidad *a priori* de la hipótesis A_n y la fórmula (6) para $\mathbb{P}(A_n|B)$ se llama la *regla de Bayes para la probabilidad a posteriori* de la hipótesis A_n .

Nota Bene. Advertimos al lector que no trate de memorizar la fórmula (6). Matemáticamente, solo se trata de una forma especial de escribir la fórmula (5) y de nada más. \square

Ejemplo 1.10 (Canal de comunicación binario). Un canal de comunicación binario simple transporta mensajes usando solo dos señales: 0 y 1. Supongamos que en un canal de comunicación binario dado el 40 % de las veces se transmite un 1; que si se transmitió un 0 la probabilidad de recibirlo correctamente es 0.90; y que si se transmitió un 1 la probabilidad de recibirlo correctamente es 0.95. Queremos determinar

- (a) la probabilidad de recibir un 1;
- (b) dado que se recibió un 1, la probabilidad de que haya sido transmitido un 1;

Solución. Consideramos los eventos $A = \text{"se transmitió un 1"}$ y $B = \text{"se recibió un 1"}$. La información dada en el enunciado del problema significa que $\mathbb{P}(A) = 0.4$, $\mathbb{P}(A^c) = 0.6$, $\mathbb{P}(B|A) = 0.95$, $\mathbb{P}(B|A^c) = 0.1$, $\mathbb{P}(B^c|A) = 0.05$, $\mathbb{P}(B^c|A^c) = 0.90$ y se puede representar en la forma de un diagrama de árbol tal como se indicó en la sección 1.2.

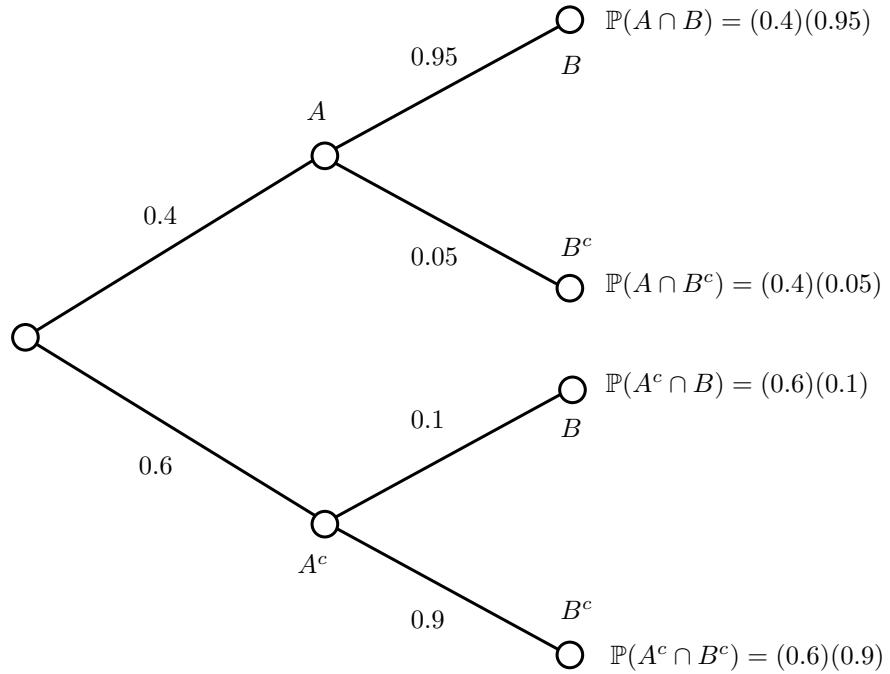


Figura 4: Observando el árbol se deduce que la probabilidad de recibir un 1 es $\mathbb{P}(B) = (0.4)(0.95) + (0.6)(0.1) = 0.44$. También se deduce que la probabilidad de que haya sido transmitido un 1 dado que se recibió un 1 es $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{(0.4)(0.95)}{0.44} = 0.863\dots$. \square

Ejercicios adicionales

1. Los dados de Efron. Se trata de cuatro dados A, B, C, D como los que se muestran en la Figura 5.

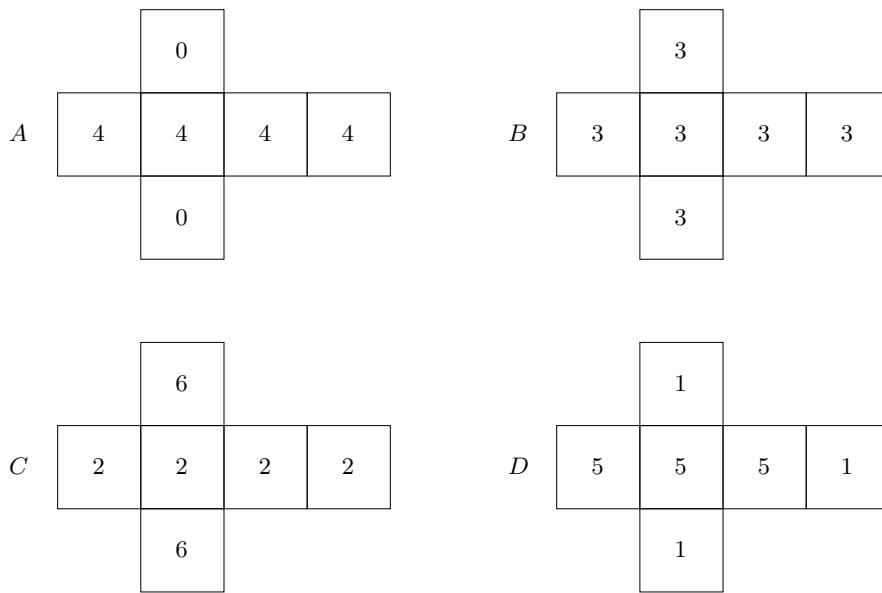


Figura 5: Dados de Efron

Las reglas del juego son las siguientes: juegan dos jugadores, cada jugador elige un dado, se tiran los dados y gana el que obtiene el número más grande.

(a) Calcular las siguientes probabilidades: que A le gane a B ; que B le gane a C ; que C le gane a D ; que D le gane a A .

(b) ¿Cuál es la mejor estrategia para jugar con los dados de Efron?.

(c) Lucas y Monk jugaran con los dados de Efron eligiendo los dados al azar. Calcular las siguientes probabilidades:

- que Lucas pierda la partida si Monk obtiene un 3,
- que Lucas gane la partida si le toca el dado A .

(d) ¿Qué ocurre con el juego cuando los dados se eligen al azar?

(e) ¿Qué ocurre con el juego si a un jugador se le permite elegir un dado y el otro debe elegir al azar uno entre los restantes tres?

(f) Lucas y Monk jugaron con los dados de Efron, eligiendo los dados al azar. Lucas ganó, ¿cuál es la probabilidad de que le haya tocado el dado C ?

2. Independencia estocástica

Definición 2.1 (Independencia estocástica). Los eventos A_1, A_2, \dots, A_n son *mutuamente independientes* si satisfacen las siguientes $2^n - n - 1$ ecuaciones:

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_m}), \quad (7)$$

donde $m = 1, 2, \dots, n$, y $1 \leq i_1 < i_2 < \dots < i_m \leq n$.

Nota Bene 1. Para $n = 2$ el sistema de ecuaciones (7) se reduce a una condición: dos eventos A_1 y A_2 son independientes si satisfacen la ecuación

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2). \quad (8)$$

Ejemplo 2.2.

(a) Se extrae un naípe al azar de un mazo de naipes de poker. Por razones de simetría esperamos que los eventos “corazón” y “As” sean independientes. En todo caso, sus probabilidades son $1/4$ y $1/13$, respectivamente y la probabilidad de su realización simultánea es $1/52$.

(b) Se arrojan dos dados. Los eventos “as en el primer dado” y “par en el segundo” son independientes pues la probabilidad de su realización simultánea, $3/36 = 1/12$, es el producto de sus probabilidades respectivas: $1/6$ y $1/2$.

(c) En una permutación aleatoria de las cuatro letras a, b, c, d los eventos “ a precede a b ” y “ c precede a d ” son independientes. Esto es intuitivamente claro y fácil de verificar. \square

Nota Bene 2. Para $n > 2$, los eventos A_1, A_2, \dots, A_n pueden ser independientes de a pares: $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$, $1 \leq i < j \leq n$, pero no ser mutuamente independientes.

Ejemplo 2.3. Sea Ω un conjunto formado por cuatro elementos: $\omega_1, \omega_2, \omega_3, \omega_4$; las correspondientes probabilidades elementales son todas iguales a $1/4$. Consideramos tres eventos:

$$A_1 = \{\omega_1, \omega_2\}, \quad A_2 = \{\omega_1, \omega_3\}, \quad A_3 = \{\omega_1, \omega_4\}.$$

Es fácil ver que los eventos A_1, A_2, A_3 son independientes de a pares, pero no son mutuamente independientes:

$$\begin{aligned} \mathbb{P}(A_1) &= \mathbb{P}(A_2) = \mathbb{P}(A_3) = 1/2, \\ \mathbb{P}(A_1 \cap A_2) &= \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_2 \cap A_3) = 1/4 = (1/2)^2, \\ \mathbb{P}(A_1 \cap A_2 \cap A_3) &= 1/4 \neq (1/2)^3. \end{aligned}$$

\square

Independencia y probabilidades condicionales. Para introducir el concepto de independencia no utilizamos probabilidades condicionales. Sin embargo, sus aplicaciones dependen generalmente de las propiedades de ciertas probabilidades condicionales.

Para fijar ideas, supongamos que $n = 2$ y que las probabilidades de los eventos A_1 y A_2 son positivas. En tal caso, los eventos A_1 y A_2 son independientes si y solamente si

$$\mathbb{P}(A_2|A_1) = \mathbb{P}(A_2) \quad \text{y} \quad \mathbb{P}(A_1|A_2) = \mathbb{P}(A_1).$$

El siguiente Teorema expresa la relación general entre el concepto de independencia y las probabilidades condicionales.

Teorema 2.4. Sean A_1, A_2, \dots, A_n eventos tales que todas las probabilidades $\mathbb{P}(A_i)$ son positivas. Una condición necesaria y suficiente para la mutua independencia de los eventos A_1, A_2, \dots, A_n es la satisfacción de las ecuaciones

$$\mathbb{P}(A_i | A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_i) \quad (9)$$

cualesquiera sean i_1, i_2, \dots, i_k, i distintos dos a dos.

Ejercicios adicionales

2. Se tira una moneda honesta n veces. Sea A el evento que se obtenga al menos una cara y sea B el evento que se obtengan al menos una cara y al menos una ceca. Analizar la independencia de los eventos A y B .

3. Andrés, Francisco, Jemina e Ignacio fueron amigos en la escuela primaria. Se reencontraron en el curso 23 (PyE 61.09) de la FIUBA y se reunieron de a parejas a charlar. Como resultado de esas charlas, cada pareja renovó su amistad con probabilidad $1/2$ y no lo hizo con probabilidad $1/2$, independientemente de las demás. Posteriormente, Andrés recibió un rumor y lo transmitió a todas sus amistades. Suponiendo que cada uno de los que reciba un rumor lo transmitirá a todas sus amistades, cuál es la probabilidad de que Ignacio haya recibido el rumor transmitido por Andrés?.

3. Modelos discretos

Los espacios muestrales más simples son aquellos que contienen un número finito, n , de puntos. Si n es pequeño (como en el caso de tirar algunas monedas), es fácil visualizar el espacio. El espacio de distribuciones de cartas de poker es más complicado. Sin embargo, podemos imaginar cada punto muestral como una ficha y considerar la colección de esas fichas como representantes del espacio muestral. Un evento A se representa por un determinado conjunto de fichas, su complemento A^c por las restantes. De aquí falta sólo un paso para imaginar una bol con infinitas fichas o un espacio muestral con una sucesión infinita de puntos $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$.

Definición 3.1. Un espacio muestral se llama discreto si contiene finitos o infinitos puntos que pueden ordenarse en una sucesión $\omega_1, \omega_2, \dots$.

Sean Ω un conjunto infinito numerable y \mathcal{A} la σ -álgebra de todos los subconjuntos contenidos en Ω . Todos los espacios de probabilidad que se pueden construir sobre (Ω, \mathcal{A}) se obtienen de la siguiente manera:

1. Tomamos una sucesión de números no negativos $\{p(\omega) : \omega \in \Omega\}$ tal que

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

2. Para cada evento $A \in \mathcal{A}$ definimos $\mathbb{P}(A)$ como la suma de las probabilidades de los eventos elementales contenidos en A :

$$\mathbb{P}(A) := \sum_{\omega \in A} p(\omega). \quad (10)$$

Nombres. La función $p : \Omega \rightarrow [0, 1]$ que asigna probabilidades a los eventos elementales $\omega \in \Omega$ se llama *función de probabilidad*. La función $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$ definida en (10) se llama *la medida de probabilidad inducida por p*.

Nota Bene 1. De la definición (10) resultan inmediatamente las siguientes propiedades

- (i) Para cada $A \in \mathcal{A}$ vale que $\mathbb{P}(A) \geq 0$
- (ii) $\mathbb{P}(\Omega) = 1$.
- (iii) σ -*aditividad*. Si A_1, A_2, \dots es una sucesión de eventos disjuntos dos a dos, entonces

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Nota Bene 2. No se excluye la posibilidad de que un punto tenga probabilidad cero. Esta convención parece artificial pero es necesaria para evitar complicaciones. En espacios discretos probabilidad cero se interpreta como imposibilidad y cualquier punto muestral del que se sabe que tiene probabilidad cero puede suprimirse impunemente del espacio muestral. Sin embargo, frecuentemente los valores numéricos de las probabilidades no se conocen de antemano, y se requieren complicadas consideraciones para decidir si un determinado punto muestral tiene o no probabilidad positiva.

Distribución geométrica

Ejemplo 3.2 (Probabilidad geométrica). Sea p un número real tal que $0 < p < 1$. Observando que

$$\sum_{n=1}^{\infty} (1-p)^{n-1} = \frac{1}{p},$$

se deduce que la función $p : \mathbb{N} \rightarrow \mathbb{R}$ definida por

$$p(n) := (1-p)^{n-1}p, \quad n = 1, 2, \dots$$

define una función de probabilidad en $\Omega = \mathbb{N} = \{1, 2, 3, \dots\}$ que se conoce por el nombre de *distribución geométrica de parámetro p*. Esta función de probabilidades está íntimamente relacionada con la cantidad de veces que debe repetirse un experimento aleatorio para que ocurra un evento A (prefijado de antemano) cuya probabilidad de ocurrencia en cada experimento individual es p . \square

Ejemplo 3.3. El experimento consiste en lanzar una moneda tantas veces como sea necesario hasta que salga cara. El resultado del experimento será la cantidad de lanzamientos necesarios hasta que se obtenga cara. Los resultados posibles son

$$\Omega = \{1, 2, 3, \dots\} \cup \{\infty\}.$$

El símbolo ∞ está puesto para representar la posibilidad de que todas las veces que se lanza la moneda el resultado obtenido es ceca. El primer problema que debemos resolver es asignar probabilidades a los puntos muestrales. Una forma de resolverlo es la siguiente. Cada vez que se arroja una moneda los resultados posibles son cara (H) o ceca (T). Sean p y q la probabilidad

de observar cara y ceca, respectivamente, en cada uno de los lanzamientos. Claramente, p y q deben ser no negativos y

$$p + q = 1.$$

Suponiendo que cada lanzamiento es independiente de los demás, las probabilidades se multiplican. En otras palabras, *la probabilidad de cada secuencia determinada es el producto obtenido de reemplazar las letras H y T por p y q, respectivamente*. Así,

$$\mathbb{P}(H) = p; \quad \mathbb{P}(TH) = qp; \quad \mathbb{P}(TTH) = qqp; \quad \mathbb{P}(TTTH) = qqqp.$$

Puede verse que para cada $n \in \mathbb{N}$ la secuencia formada por $n - 1$ letras T seguida de la letra H debe tener probabilidad $q^{n-1}p = (1 - p)^{n-1}p$.

El argumento anterior sugiere la siguiente asignación de probabilidades sobre Ω : para cada $n \in \mathbb{N}$, $p(n)$, la probabilidad de que la primera vez que se obtiene cara ocurra en el n -ésimo lanzamiento de la moneda está dada por

$$p(n) = (1 - p)^{n-1}p.$$

Como las probabilidades geométricas suman 1 (ver el ejemplo 3.2) al resultado “ceca en todos los tiros” se le debe asignar probabilidad $p(\infty) = 0$. Como el espacio muestral es discreto no hay problema en suprimir el punto ∞ .

Consideremos el evento $A =$ “se necesitan una cantidad par de tiros para obtener la primer cara”. Entonces,

$$A = \{2, 4, 6, 8, \dots\},$$

y

$$\begin{aligned} \mathbb{P}(A) &= \sum_{\omega \in A} p(\omega) = \sum_{k=1}^{\infty} p(2k) = \sum_{k=1}^{\infty} q^{2k-1}p = pq \sum_{k=0}^{\infty} q^{2k} = pq \left(\frac{1}{1 - q^2} \right) \\ &= \frac{pq}{(1 - q)(1 + q)} = \frac{q}{1 + q} = \frac{1 - p}{2 - p}. \end{aligned}$$

□

Ejemplo 3.4. Lucas y Monk juegan a la moneda. Lanzan una moneda equilibrada al aire, si sale cara, Lucas le gana un peso a Monk; si sale ceca, Monk le gana un peso a Lucas. El juego termina cuando alguno gana dos veces seguidas.

El espacio muestral asociado a este experimento aleatorio es

$$\Omega = \{HH, TT, HTT, THH, HTHH, THTT, \dots\}.$$

Como podemos tener secuencias de cualquier longitud de caras y cecas alternadas, el espacio muestral es necesariamente infinito.

El evento $A_1 =$ “la moneda fue lanzada como máximo tres veces” está dado por todos los elementos de Ω que tienen longitud menor o igual que tres:

$$A_1 = \{HH, TT, HTT, THH\}$$

y su probabilidad es

$$\mathbb{P}(A_1) = \mathbb{P}(HH) + \mathbb{P}(TT) + \mathbb{P}(HTT) + \mathbb{P}(THH) = \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} = \frac{3}{4}.$$

El evento $A_2 = \text{"ceca en el primer lanzamiento"}$ está dado por todos los elementos de Ω que comienzan con T :

$$A_2 = \{TT, THH, THTT, THTHH, \dots\},$$

y su probabilidad es

$$\begin{aligned}\mathbb{P}(A_2) &= \mathbb{P}(TT) + \mathbb{P}(THH) + \mathbb{P}(THTT) + \mathbb{P}(THTHH) + \dots \\ &= \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^5} + \dots = \frac{1}{2}.\end{aligned}$$

¿Cuál es la probabilidad de que el juego termine alguna vez? Si definimos los eventos $A_n := \text{"el juego termina en la } n\text{-ésima jugada"}$, $n \geq 2$, tendremos que el evento “el juego termina alguna vez” es la unión disjunta de los eventos A_1, A_2, \dots , y por lo tanto su probabilidad es la suma de las probabilidades de los eventos A_n . Para cada $n \geq 2$ la probabilidad de A_n es

$$\mathbb{P}(A_n) = \frac{2}{2^n} = \frac{1}{2^{n-1}}$$

En consecuencia la probabilidad de que el juego termine alguna vez es

$$\sum_{n \geq 2} \frac{1}{2^{n-1}} = \sum_{n \geq 1} \frac{1}{2^n} = 1.$$

□

Distribución de Poisson

Ejemplo 3.5 (Probabilidad de Poisson). Sea λ un número real positivo. Observando que

$$e^\lambda = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!},$$

se deduce que la función $p : \mathbb{N}_0 \rightarrow \mathbb{R}$ definida por

$$p(n) := e^{-\lambda} \frac{\lambda^n}{n!}, \quad n = 0, 1, 2, \dots$$

define una función de probabilidad en $\Omega = \mathbb{N}_0 = \{0, 1, 2, \dots\}$, conocida como *la distribución de Poisson de intensidad λ* .

□

4. Modelos continuos

4.1. Puntos al azar sobre un segmento. La distribución uniforme

Elegir un punto al azar dentro de un segmento de recta de longitud finita es un experimento conceptual intuitivamente claro. Desde el punto de vista teórico el experimento debe describirse mediante un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$.

No se pierde generalidad, si se supone que la longitud del segmento es la unidad y se lo identifica con el intervalo $\Omega = [0, 1]$. La σ -álgebra de eventos \mathcal{A} y la medida de probabilidad $\mathbb{P} : \mathcal{A} \rightarrow \mathbb{R}$ se construyen por etapas.

1. Definimos \mathcal{A}_0 como la familia de los intervalos contenidos en Ω de la forma $[a, b]$, $[a, b)$, $(a, b]$ o (a, b) , $a \leq b$ (notar que \mathcal{A}_0 no es un álgebra) y definimos $\mathbb{P}_0 : \mathcal{A}_0 \rightarrow \mathbb{R}$ de la siguiente manera:

$$\mathbb{P}_0(A) := \text{longitud}(A) = b - a, \text{ si los extremos del intervalo } A \text{ son } a \text{ y } b.$$

2. La familia \mathcal{A}_1 de todas las uniones finitas de conjuntos disjuntos de \mathcal{A}_0 es un álgebra de eventos y la función $\mathbb{P}_1 : \mathcal{A}_1 \rightarrow \mathbb{R}$ definida por

$$\mathbb{P}_1(A) := \sum_{i=1}^k \mathbb{P}_0(A_i), \text{ si } A = \bigcup_{i=1}^k A_i,$$

donde $A_1, \dots, A_k \in \mathcal{A}_0$ y $A_i \cap A_j = \emptyset$ para toda pareja de índices $i \neq j$, es una medida de probabilidad (pues satisface los axiomas I-IV).

3. El teorema de extensión se ocupa del resto: la medida de probabilidad \mathbb{P}_1 definida sobre el álgebra \mathcal{A}_1 se extiende únicamente a una medida de probabilidad \mathbb{P} definida sobre la σ -álgebra generada por \mathcal{A}_1 , $\mathcal{A} := \sigma(\mathcal{A}_1)$.

Nota Bene. Esta definición de probabilidad que a cada intervalo $A \subset [0, 1]$ le asigna su respectiva longitud se llama la *distribución uniforme sobre el intervalo* $[0, 1]$ y constituye una generalización de la noción de equiprobabilidad sobre la que se basa la definición de Laplace de la probabilidad para espacios finitos: “*casos favorables sobre casos posibles*”.

4.2. Geometría y probabilidad

Una construcción completamente análoga a la de la sección anterior permite describir teóricamente el experimento conceptual, intuitivamente claro, que consiste en *elegir un punto al azar dentro de una región plana*, $\Lambda \subset \mathbb{R}^2$, *de área finita y no nula*. Para fijar ideas, se puede imaginar que la región plana es un blanco sobre el que se arroja un dardo.

Ejemplo 4.1 (Dardos). El juego de dardos consiste en tirar un dardo contra un blanco circular. Supongamos que disparamos un dardo (que acertamos al blanco) y observamos dónde se clavó. Naturalmente, los resultados posibles de este experimento son todos los puntos del blanco. No se pierde generalidad si se supone que el centro del blanco es el origen de \mathbb{R}^2 y que su radio es 1. En tal caso el espacio muestral de este experimento es

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}.$$

Intuitivamente, la probabilidad de acertarle a un punto predeterminado (arbitrario) debería ser cero. Sin embargo, la probabilidad de que el dardo se clave en cualquier subconjunto (“gordo”) A del blanco debería ser proporcional a su área y determinarse por la fracción del área del blanco contenida en A . En consecuencia, definimos

$$\mathbb{P}(A) := \frac{\text{área de } A}{\text{área del blanco}} = \frac{\text{área de } A}{\pi}.$$

Por ejemplo, si $A = \{(x, y) : x^2 + y^2 \leq r^2\}$ es el evento que el dardo caiga a distancia $r < 1$ del centro del blanco, entonces

$$\mathbb{P}(A) = \frac{\pi r^2}{\pi} = r^2.$$

□

“Puntos al azar en regiones planas”. Si hacemos abstracción de la forma circular del blanco y de la semántica involucrada en el juego de dardos, obtenemos un modelo probabilístico para el experimento conceptual que consiste en “sortear” o *elegir un punto al azar* en una región plana $\Lambda \subset \mathbb{R}^2$ de área finita y positiva. El espacio muestral es la región plana, $\Omega = \Lambda$, la σ -álgebra de los eventos, \mathcal{A} , es la familia de todos los subconjuntos de Λ a los que se les puede medir el área y la probabilidad de cada evento A es la fracción del área de Λ contenida en A . Esto es,

$$\mathbb{P}(A) := \frac{\text{área}(A)}{\text{área}(\Lambda)}. \quad (11)$$

Esta forma de asignar probabilidades es la equivalente para el caso continuo de la fórmula *casos favorables sobre casos posibles* utilizada en espacios muestrales finitos para modelar experimentos aleatorios con resultados equiprobables.

Nota Bene. Si en lugar de elegir un punto al azar dentro del segmento $[a, b]$ elegimos dos puntos de manera independiente, el experimento tendrá por resultado un par de números reales contenidos en $[a, b]$. El espacio muestral será el cuadrado de lado $[a, b]$, $\Omega = [a, b] \times [a, b]$. En este espacio la asignación de probabilidades definida en (11) resulta consistente con la noción de independencia.

Ejemplo 4.2. Se eligen al azar (y en forma independiente) dos puntos x_1 y x_2 dentro de un segmento de longitud L . Hallar la probabilidad de que la longitud del segmento limitado por los puntos x_1 y x_2 resulte menor que $L/2$.

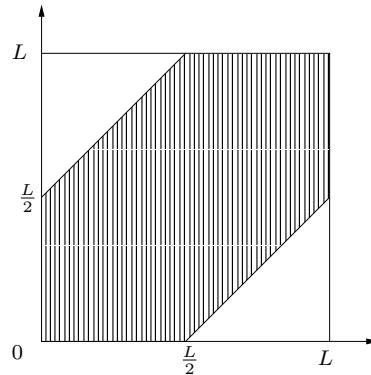


Figura 6: La región sombreada corresponde al evento A =“la longitud del segmento limitado por los puntos x_1 y x_2 resulte menor que $L/2$ ”.

El espacio muestral de este experimento es un cuadrado de lado L que puede representarse en la forma $\Omega = \{(x_1, x_2) : 0 \leq x_1 \leq L, 0 \leq x_2 \leq L\}$.

El evento A =“la longitud del segmento limitado por los puntos x_1 y x_2 resulte menor que $L/2$ ” puede ocurrir de dos maneras distintas:

- (1) si $x_1 \leq x_2$, se debe cumplir la desigualdad $x_2 - x_1 < L/2$;
- (2) si $x_2 < x_1$, debe cumplirse la desigualdad $x_1 - x_2 < L/2$.

Observando la Figura 6 está claro que el área del evento A se obtiene restando al área del cuadrado de lado L el área del cuadrado de lado $L/2$:

$$\text{área de } A = L^2 - \frac{L^2}{4} = \frac{3}{4}L^2.$$

Como el área total del espacio muestral es L^2 , resulta que $\mathbb{P}(A) = 3/4$. \square

Ejemplo 4.3 (Las agujas de Buffon). Una aguja de longitud $2l$ se arroja sobre un plano dividido por rectas paralelas. La distancia entre rectas es $2a$. Suponiendo que $l < a$, cuál es la probabilidad de que la aguja intersecte alguna de las rectas?

Localizamos la aguja mediante la distancia ρ de su centro a la recta más cercana y el ángulo agudo θ entre la recta y la aguja: $0 \leq \rho \leq a$ y $0 \leq \theta \leq \pi/2$. El rectángulo determinado por esas desigualdades es el espacio muestral Ω . El evento $A = “la\ aguja\ interseca\ la\ recta”$ ocurre si $\rho \leq l \sin \theta$. La probabilidad de A es el cociente del área de la figura determinada por las tres desigualdades $0 \leq \rho \leq a$, $0 \leq \theta \leq \pi/2$ y $\rho \leq l \sin \theta$ y el área del rectángulo $\pi a/2$.

El área de la figura es $\int_0^{\pi/2} l \sin(\theta) d\theta = l$. Por lo tanto, la probabilidad de intersección es

$$\mathbb{P}(A) = \frac{2l}{\pi a}. \quad (12)$$

La fórmula (12) indica un método aleatorio para estimar π : arrojar la aguja n veces sobre el plano y contar $n(A)$ la cantidad de veces que la aguja interesectó alguna recta:

$$\hat{\pi} = 2(l/a)(n/n(A)).$$

\square

4.3. Paradoja de Bertrand

Se dibuja una cuerda aleatoria CD sobre el círculo de radio 1. ¿Cuál es la probabilidad que la longitud de la cuerda CD supere $\sqrt{3}$, la longitud del lado del triángulo equilátero inscripto en dicho círculo?

Este es un ejemplo de un problema planteado de manera incompleta. La pregunta que debe formularse es la siguiente: ¿qué significa elegir “aleatoriamente”? Bertrand propuso tres respuestas diferentes a esa pregunta. Las diferentes respuestas corresponden en realidad a diferentes modelos probabilísticos, i.e., diferentes espacios de probabilidad concretos $(\Omega, \mathcal{A}, \mathbb{P})$.

- *Primer modelo.* Sea Ω_1 la bola de radio 1, $\Omega_1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$, con la σ -álgebra \mathcal{A} de los “subconjuntos cuya área está definida”. Para cada $A \in \mathcal{A}$,

$$\mathbb{P}_1(A) = \frac{\text{área}(A)}{\text{área}(\Omega)} = \frac{\text{área}(A)}{\pi}.$$

C y D se construyen del siguiente modo: usando la ley de distribución \mathbb{P}_1 se sortea un punto ω sobre la bola de radio 1 y CD es perpendicular al segmento $\overline{0\omega}$ cuyos extremos son $(0, 0)$ y ω . La longitud de CD es una función de ω que llamaremos $\ell(\omega)$. Queremos calcular $\mathbb{P}_1(\ell(\omega) \geq \sqrt{3})$. Notar que

$$\ell(\omega) \geq \sqrt{3} \iff \text{longitud}(\overline{0\omega}) \geq \frac{1}{2}.$$

Por lo tanto,

$$\mathbb{P}_1(\ell(\omega) \geq \sqrt{3}) = \frac{\pi - \pi/4}{\pi} = \frac{3}{4}.$$

- *Segundo modelo.* Sea Ω_2 el círculo de radio 1, $\Omega_2 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$, con la σ -álgebra \mathcal{A} de los “subconjuntos cuya longitud está definida”. Para cada $A \in \mathcal{A}$,

$$\mathbb{P}_2(A) = \frac{\text{longitud}(A)}{\text{longitud}(\Omega)} = \frac{\text{longitud}(A)}{2\pi}.$$

C y D se construyen del siguiente modo: Se fija el punto C ; con la ley \mathbb{P}_2 se sortea un punto ω sobre el círculo de radio 1 y se pone $D = \omega$. La longitud de CD es una función de ω que llamaremos $\ell(\omega)$. El conjunto $\{\omega : \ell(\omega) \geq \sqrt{3}\}$ es el segmento del círculo determinado dos vértices del triángulo equilátero inscripto en el círculo, a saber: los del lado opuesto al vértice C . Por lo tanto,

$$\mathbb{P}_2(\ell(\omega) \geq \sqrt{3}) = \frac{2\pi/3}{2\pi} = \frac{1}{3}.$$

- *Tercer modelo.* Sea Ω_3 el intervalo $[0, 1]$ con la σ -álgebra \mathcal{A} de los “subconjuntos cuya longitud está definida”. Para cada $A \in \mathcal{A}$,

$$\mathbb{P}_3(A) = \text{longitud}(A).$$

C y D se construyen del siguiente modo: se sortea un punto ω sobre el intervalo $[0, 1]$ del eje x y CD es la cuerda perpendicular al eje x que pasa por ω . Es claro que,

$$\ell(\omega) \geq \sqrt{3} \iff \omega \in [1/2, 1].$$

Por lo tanto, la tercera respuesta es $1/2$.

Nota Bene. Obtuvimos 3 respuestas diferentes: $1/4, 1/3$ y $1/2$. Sin embargo, no hay porque sorprenderse debido a que los modelos probabilísticos correspondientes a cada respuesta son diferentes. Cuál de los tres es el “bueno” es otro problema. El modelo correcto depende del mecanismo usado para dibujar la cuerda al azar. Los tres mecanismos anteriores son puramente intelectuales, y muy probablemente, no corresponden a ningún mecanismo físico. Para discriminar entre modelos probabilísticos en competencia se debe recurrir al análisis estadístico que esencialmente se basa en dos resultados de la Teoría de Probabilidad: la ley fuerte de los grandes números y el teorema central del límite. □

4.4. De las masas puntuales a la masa continua

Para concluir esta sección mostraremos un par de métodos para construir medidas de probabilidad sobre \mathbb{R}^n .

Masas puntuales. Tomamos una sucesión de puntos $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ en \mathbb{R}^n y una sucesión de números no negativos $\{p(\mathbf{x}_1), p(\mathbf{x}_2), \dots\}$ tales que

$$\sum_{i=1}^{\infty} p(\mathbf{x}_i) = 1$$

y para cada $A \subset \mathbb{R}^n$ definimos $\mathbb{P}(A)$ como la suma de las “masas puntuales”, $p(\mathbf{x}_i)$, de los puntos \mathbf{x}_i contenidos en A :

$$\mathbb{P}(A) := \sum_{\mathbf{x}_i \in A} p(\mathbf{x}_i).$$

Nota Bene. El método de las *masas puntuales* puede generalizarse de la siguiente forma: la suma $\sum_{\mathbf{x}_i}$ se reemplaza por la integral $\int d\mathbf{x}$ y las masas puntuales $p(\mathbf{x}_i)$ por una función $\rho(\mathbf{x})$ denominada *densidad de probabilidades*. Esta metodología es de uso común en mecánica: primero se consideran sistemas con masas puntuales discretas donde cada punto tiene masa finita y después se pasa a la noción de distribución de masa continua, donde cada punto tiene masa cero. En el primer caso, la masa total del sistema se obtiene simplemente sumando las masas de los puntos individuales; en el segundo caso, las masas se calculan mediante integración sobre densidades de masa. Salvo por las herramientas técnicas requeridas, no hay diferencias esenciales entre ambos casos. \square

Definición 4.4. Una *densidad de probabilidades sobre \mathbb{R}^n* es una función (“más o menos razonable”) no negativa $\rho : \mathbb{R}^n \rightarrow \mathbb{R}^+$ tal que

$$\int_{\mathbb{R}^n} \rho(\mathbf{x}) d\mathbf{x} = 1.$$

Masa continua. Tomamos una densidad de probabilidades $\rho : \mathbb{R}^n \rightarrow \mathbb{R}^+$ y para cada subconjunto $A \subset \mathbb{R}^n$ (“más o menos razonable”) y definimos $\mathbb{P}(A)$ como la integral de la densidad $\rho(\mathbf{x})$ sobre el conjunto A :

$$\mathbb{P}(A) := \int_A \rho(\mathbf{x}) d\mathbf{x}$$

Ejemplo 4.5 (Gaussiana). La función $\rho : \mathbb{R}^2 \rightarrow \mathbb{R}^+$ definida por

$$\rho(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right)$$

es una densidad de probabilidades sobre \mathbb{R}^2 denominada *gaussiana bidimensional*. En efecto,

$$\begin{aligned} \iint_{\mathbb{R}^2} 2\pi\rho(x, y) dx dy &= \iint_{\mathbb{R}^2} \exp\left(-\frac{x^2 + y^2}{2}\right) dx dy \\ &= 2 \iint_{\mathbb{R}^2} \exp(-(x^2 + y^2)) dx dy \\ &= 2 \int_0^{2\pi} \left(\int_0^\infty e^{-\rho^2} \rho d\rho \right) d\theta \\ &= \int_0^{2\pi} \left(\int_0^\infty e^{-\rho^2} 2\rho d\rho \right) d\theta \\ &= 2\pi. \end{aligned} \tag{13}$$

Nota Bene. Observando con cuidado las identidades (13) se puede ver que

$$\int_{\mathbb{R}} e^{-x^2/2} dx = \sqrt{2\pi}.$$

Por lo tanto, la función $\varphi : \mathbb{R} \rightarrow \mathbb{R}^+$ definida por

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

es una densidad de probabilidades sobre \mathbb{R} . \square

5. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Bertsekas, D. P., Tsitsiklis, J. N.: Introduction to Probability. M.I.T. Lecture Notes. (2000)
2. Brémaud, P.: An Introduction to Probabilistic Modeling. Springer, New York. (1997)
3. Durrett, R. Elementary Probability for Applications. Cambridge University Press, New York. (2009)
4. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons, New York. (1957)
5. Grinstead, C. M. & Snell, J. L. Introduction to Probability. American Mathematical Society. (1997)
6. Meester, R.: A Natural Introduction to Probability Theory. Birkhauser, Berlin. (2008)
7. Meyer, P. L.: Introductory Probability and Statistical Applications. Addison-Wesley, Massachusetts. (1972)
8. Ross, S. M: Introduction to Probability and Statistics for Engineers and Scientists. Elsevier Academic Press, San Diego. (2004)
9. Skorokhod, A. V.: Basic Principles and Applications of Probability Theory. Springer-Verlag, Berlin. (2005)
10. Soong, T. T.: Fundamentals of Probability and Statistics for Engineers. John Wiley & Sons Ltd. (2004)

Variables aleatorias: nociones básicas (Borradores, Curso 23)

Sebastian Grynberg

20 de marzo 2013



... el único héroe válido es el héroe “en grupo”,
nunca el héroe individual, el héroe solo.

(Héctor G. Oesterheld)

Índice

1. Variables aleatorias	3
1.1. Propiedades de la función de distribución	6
1.2. Clasificación de variables aleatorias	7
1.3. Cuantiles	11
1.4. Construcción de variables aleatorias	13
1.5. Función de distribución empírica e histogramas	17
2. Variables truncadas	21
2.1. Perdida de memoria	22
2.2. Caracterización cualitativa de la distribución exponencial	23
2.3. Dividir y conquistar	23
3. Bibliografía consultada	24

1. Variables aleatorias

Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad. Una *variable aleatoria* sobre Ω es una función $X : \Omega \rightarrow \mathbb{R}$ tal que para todo $x \in \mathbb{R}$

$$\{X \leq x\} := \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A},$$

i.e., para todo $x \in \mathbb{R}$ el evento $\{X \leq x\}$ tiene asignada probabilidad. La *función de distribución* $F_X : \mathbb{R} \rightarrow [0, 1]$ de la variable aleatoria X se define por

$$F_X(x) := \mathbb{P}(X \leq x).$$

Cálculo de probabilidades. La función de distribución resume (y contiene) toda la información relevante sobre de la variable aleatoria. Para ser más precisos, para cada pareja de números reales $a < b$ vale que ¹

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a). \quad (1)$$

□

Ejemplos

Ejemplo 1.1 (Dado equilibrado). Sea X el resultado del lanzamiento de un dado equilibrado. Los posibles valores de X son $1, 2, 3, 4, 5, 6$. Para cada $k \in \{1, 2, 3, 4, 5, 6\}$ la probabilidad de que X tome el valor k es $1/6$.

Sea $x \in \mathbb{R}$. Si $x < 1$ es evidente que $\mathbb{P}(X \leq x) = 0$. Si $k \leq x < k + 1$ para algún $k \in \{1, 2, 3, 4, 5\}$ la probabilidad del evento $\{X \leq x\}$ es la probabilidad de observar un valor menor o igual que k y en consecuencia, $\mathbb{P}(X \leq x) = k/6$. Finalmente, si $x \geq 6$ es evidente que $\mathbb{P}(X \leq x) = 1$.

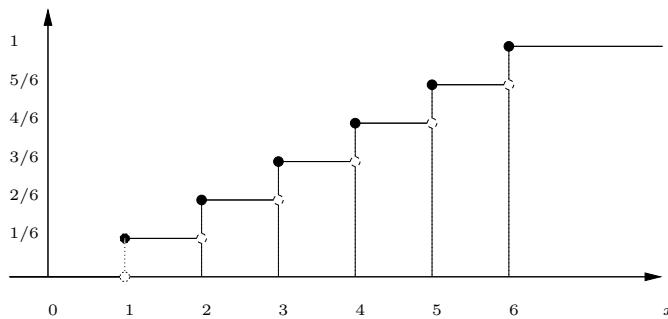


Figura 1: Gráfico de la función de distribución del resultado de lanzar un dado equilibrado.

Por lo tanto, la función de distribución de X se puede expresar del siguiente modo

$$F_X(x) = \sum_{k=1}^6 \frac{1}{6} \mathbf{1}\{k \leq x\}.$$

□

¹Basta observar que $\{X \leq a\} \subset \{X \leq b\}$ y usar las propiedades de la probabilidad. De la igualdad $\{a < X \leq b\} = \{X \leq b\} \setminus \{X \leq a\}$ se deduce que $\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a)$.

Ejemplo 1.2 (Fiabilidad). Un problema fundamental de la ingeniería es el problema de la *fiabilidad*. Informalmente, la fiabilidad de un sistema se define como su capacidad para cumplir ciertas funciones prefijadas. Esta propiedad se conserva durante un período de tiempo hasta que ocurre una *falla* que altera la capacidad de trabajo del sistema. Por ejemplo: rupturas y cortocircuitos; fracturas, deformaciones y atascamientos de piezas mecánicas; el fundido o la combustión de las componentes de un circuito.

Debido a que las fallas pueden ocurrir como hechos casuales, podemos considerar que *el tiempo de funcionamiento, T , hasta la aparición de la primer falla* es una variable aleatoria a valores no negativos.

La fiabilidad de un sistema se caracteriza por su *función intensidad de fallas $\lambda(t)$* . Esta función temporal tiene la siguiente propiedad: cuando se la multiplica por dt se obtiene la probabilidad condicional de que el sistema sufra una falla durante el intervalo de tiempo $(t, t + dt]$ sabiendo que hasta el momento t funcionaba normalmente. Si se conoce la función $\lambda(t)$ se puede hallar la ley de distribución de probabilidades de T .

Para calcular la función de distribución de T estudiaremos dos eventos: $A := \{T > t\}$ (el sistema funciona hasta el momento t) y $B := \{t < T \leq t + dt\}$ (el sistema sufre una falla en el intervalo de tiempo $(t, t + dt]$). Como $B \subset A$, tenemos que $\mathbb{P}(B) = \mathbb{P}(B \cap A)$ y de la regla del producto se deduce que

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A). \quad (2)$$

Si la función de distribución de T admite derivada continua, salvo términos de segundo orden que se pueden despreciar, la probabilidad del evento B se puede expresar en la forma

$$\mathbb{P}(B) = \mathbb{P}(t < T \leq t + dt) = F_T(t + dt) - F_T(t) = F'_T(t)dt. \quad (3)$$

La probabilidad del evento A se puede expresar en la forma

$$\mathbb{P}(A) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F_T(t). \quad (4)$$

Finalmente, la probabilidad condicional $\mathbb{P}(B|A)$ se expresa mediante la función intensidad de fallas $\lambda(t)$:

$$\mathbb{P}(B|A) = \lambda(t)dt \quad (5)$$

Sustituyendo las expresiones (3)-(5) en la fórmula (2) obtenemos, después de dividir ambos miembros por dt , una ecuación diferencial de primer orden para $F_T(t)$

$$F'_T(t) = \lambda(t)(1 - F_T(t)). \quad (6)$$

Debido a que la duración del servicio del sistema no puede ser negativa, el evento $\{T \leq 0\}$ es imposible. En consecuencia, $F_T(0) = 0$. Integrando la ecuación diferencial (6) con la condición inicial $F(0) = 0$, obtenemos ²

$$F_T(t) = 1 - \exp \left(- \int_0^t \lambda(s)ds \right). \quad (7)$$

²

$F'_T(t) = \lambda(t)(1 - F_T(t)) \iff \frac{F'_T(t)}{1 - F_T(t)} = \lambda(t) \iff \frac{d}{dt} \log(1 - F_T(t)) = -\lambda(t)$
 $\iff \log(1 - F_T(t)) = - \int_0^t \lambda(s)ds + C \iff F_T(t) = 1 - \exp \left(- \int_0^t \lambda(s)ds + C \right).$

Usando que $F_T(0) = 0$ se deduce que $C = 0$.

Nota Bene. El desarrollo anterior presupone que la función intensidad de fallas $\lambda(t)$ verifica las siguientes condiciones: (1) $\lambda(t) \geq 0$ para todo $t > 0$ y (2) $\int_0^\infty \lambda(t)dt = +\infty$. \square

Ejemplo 1.3 (Fiabilidad). Se estipula que la duración de servicio de un sistema automático debe ser t_0 . Si durante ese período el sistema falla, se lo repara y se lo utiliza hasta que sirva el plazo estipulado. Sea S el tiempo de funcionamiento del sistema después de la primera reparación. Queremos hallar la función de distribución de S .

En primer lugar observamos que la relación entre la variable aleatoria S y el instante T en que ocurre la primera falla del sistema es la siguiente

$$S = \max(t_0 - T, 0) = \begin{cases} t_0 - T & \text{si } T \leq t_0, \\ 0 & \text{si } T > t_0. \end{cases}$$

Sea $F_S(s)$ la función de distribución de la variable S . Es claro que para $s < 0$, $F_S(s) = 0$ y que para $s \geq t_0$, $F_S(s) = 1$. Lo que falta hacer es analizar el comportamiento de F_S sobre el intervalo $0 \leq s < t_0$. Sea $s \in [0, t_0]$

$$\begin{aligned} F_S(s) &= \mathbb{P}(S \leq s) = \mathbb{P}(\max(t_0 - T, 0) \leq s) = \mathbb{P}(t_0 - T \leq s, 0 \leq s) \\ &= \mathbb{P}(t_0 - T \leq s) = \mathbb{P}(t_0 - s \leq T) = \exp\left(-\int_0^{t_0-s} \lambda(t)dt\right), \end{aligned}$$

donde $\lambda(t)$ es la función intensidad de fallas del sistema.

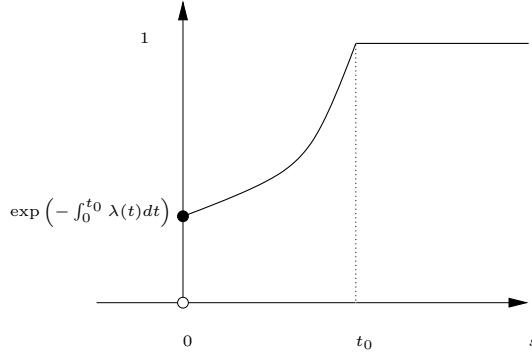


Figura 2: Gráfico de la función de distribución de la variable aleatoria S .

Por lo tanto,

$$F_S(s) = \exp\left(-\int_0^{t_0-s} \lambda(t)dt\right) \mathbf{1}\{0 \leq s < t_0\} + \mathbf{1}\{s \geq t_0\}.$$

\square

Ejercicios adicionales

1. Sea X una variable aleatoria con función de distribución $F_X(x)$. Mostrar que para cada pareja de números reales $a < b$ vale que:

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) + \mathbb{P}(X = a) \quad (8)$$

$$\mathbb{P}(a \leq X < b) = F_X(b) - \mathbb{P}(X = b) - F_X(a) + \mathbb{P}(X = a) \quad (9)$$

$$\mathbb{P}(a < X < b) = F_X(b) - \mathbb{P}(X = b) - F_X(a) \quad (10)$$

Notar que las fórmulas (8)-(10), junto con (1), muestran como calcular la probabilidad de que la variable aleatoria X tome valores en un intervalo de extremos a y b y contienen una advertencia sobre la acumulación de masa positiva en alguno de los dos extremos.

1.1. Propiedades de la función de distribución

Lema 1.4. Sea $X : \Omega \rightarrow \mathbb{R}$ una variable aleatoria. La función de distribución de X , $F_X(x) = \mathbb{P}(X \leq x)$, tiene las siguientes propiedades:

- (F1) es *no decreciente*: si $x_1 \leq x_2$, entonces $F_X(x_1) \leq F_X(x_2)$;
- (F2) es *continua a derecha*: para todo $x_0 \in \mathbb{R}$ vale que $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$;
- (F3) $\lim_{x \rightarrow -\infty} F_X(x) = 0$ y $\lim_{x \rightarrow \infty} F_X(x) = 1$.

Demostración.

La propiedad (F1) se deduce de la fórmula (1).

La propiedad (F2) es consecuencia del axioma de continuidad de la medida de probabilidad \mathbb{P} . Se considera una sucesión decreciente de números positivos que converge a 0, $\epsilon_1 > \epsilon_2 > \dots > 0$, arbitraria, pero fija y se definen eventos $A_n = \{x_0 < X \leq x_0 + \epsilon_n\}$. Se observa que $A_1 \supset A_2 \supset \dots$ y $\bigcap_{n \in \mathbb{N}} A_n = \emptyset$:

$$0 = \lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \lim_{n \rightarrow \infty} \mathbb{P}(x_0 < X \leq x_0 + \epsilon_n) = \lim_{n \rightarrow \infty} F(x_0 + \epsilon_n) - F(x_0).$$

Por lo tanto,

$$F(x_0) = \lim_{n \rightarrow \infty} F(x_0 + \epsilon_n).$$

Las propiedades (F3) se demuestran de manera similar. □

Observación 1.5. Si se define

$$F_X(x_0^-) := \lim_{x \uparrow x_0} F_X(x),$$

entonces $F_X(x_0^-) = \mathbb{P}(X < x_0)$. Por lo tanto, $\mathbb{P}(X = x_0) = F_X(x_0) - F_X(x_0^-)$. En particular, si $F_X(x)$ es continua en x_0 , entonces $\mathbb{P}(X = x_0) = 0$. Si $\mathbb{P}(X = x_0) > 0$, entonces $F_X(x)$ es discontinua en x_0 y su discontinuidad es un salto de altura $\mathbb{P}(X = x_0) > 0$.

Ejercicios adicionales

2. Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad y $X : \Omega \rightarrow \mathbb{R}$ una variable aleatoria con función de distribución $F_X(x)$.

(a) Mostrar que

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad \text{y} \quad \lim_{x \rightarrow \infty} F_X(x) = 1.$$

(*Sugerencia.* Considerar sucesiones de eventos $B_n = \{X \leq -n\}$ y $C_n = \{X \leq n\}$, $n \in \mathbb{N}$, y utilizar el axioma de continuidad de la medida de probabilidad \mathbb{P} .)

(b) Mostrar que

$$\lim_{x \uparrow x_0} F_X(x) = \mathbb{P}(X < x_0).$$

(*Sugerencia.* Observar que si $x \uparrow x_0$, entonces $\{X \leq x\} \uparrow \{X < x_0\}$ y utilizar el axioma de continuidad de la medida de probabilidad \mathbb{P} .)

1.2. Clasificación de variables aleatorias

En todo lo que sigue, X designa una variable aleatoria definida sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ y $F_X(x) := \mathbb{P}(X \leq x)$ su función de distribución.

Nota Bene. Al observar el gráfico de una función de distribución lo primero que llama la atención son sus saltos y sus escalones.

Átomos. Diremos que $a \in \mathbb{R}$ es un *átomo* de $F_X(x)$ si su peso es positivo: $\mathbb{P}(X = a) = F_X(a) - F_X(a-) > 0$.

El conjunto de todos los átomos de $F_X(x)$: $\mathbb{A} = \{a \in \mathbb{R} : F_X(a) - F_X(a-) > 0\}$, coincide con el conjunto de todos los puntos de discontinuidad de $F_X(x)$. El peso de cada átomo coincide con la longitud del salto dado por la función de distribución en dicho átomo. En consecuencia, existen a lo sumo un átomo de probabilidad $> \frac{1}{2}$, a lo sumo dos átomos de probabilidad $> \frac{1}{3}$, etcétera. Por lo tanto, es posible reordenar los átomos en una sucesión a_1, a_2, \dots tal que $\mathbb{P}(X = a_1) \geq \mathbb{P}(X = a_2) \geq \dots$. En otras palabras, *existen a lo sumo numerables átomos*.

La propiedad de σ -aditividad de la medida de probabilidad \mathbb{P} implica que el peso total del conjunto \mathbb{A} no puede exceder la unidad: $\sum_{a \in \mathbb{A}} \mathbb{P}(X = a) \leq 1$.

Definición 1.6 (Variables discretas). Diremos que X es una variable aleatoria *discreta* si

$$\sum_{a \in \mathbb{A}} \mathbb{P}(X = a) = 1.$$

En tal caso, la función $p_X : \mathbb{A} \rightarrow \mathbb{R}$ definida por $p_X(x) = \mathbb{P}(X = x)$ se denomina la *función de probabilidad* de X .

Escalones. Sea X una variable aleatoria discreta. Si $a_1 < a_2$ son dos átomos consecutivos, entonces $F_X(x) = F_X(a_1)$ para todo $x \in (a_1, a_2)$. En otras palabras, *la función de distribución de una variable aleatoria discreta debe ser constante entre saltos consecutivos*.

Si no lo fuera, deberían existir dos números $x_1 < x_2$ contenidos en el intervalo (a_1, a_2) tales que $F_X(x_1) < F_X(x_2)$. En tal caso,

$$\begin{aligned} \mathbb{P}(X \in \mathbb{A} \cup (x_1, x_2]) &= \mathbb{P}(X \in \mathbb{A}) + \mathbb{P}(x_1 < X \leq x_2) = \sum_{a \in \mathbb{A}} \mathbb{P}(X = a) + F_X(x_2) - F_X(x_1) \\ &= 1 + F_X(x_2) - F_X(x_1) > 1. \end{aligned}$$

lo que constituye un absurdo. □

Definición 1.7 (Variables continuas). Diremos que X es una variable aleatoria *continua* si su función de distribución es continua.

Definición 1.8 (Variables mixtas). Diremos que X es una variable aleatoria *mixta* si no es continua ni discreta.

Definición 1.9 (Variables absolutamente continuas). Diremos que X es *absolutamente continua* si existe una función (medible) $f_X : \mathbb{R} \rightarrow \mathbb{R}^+$, llamada *densidad* de X , tal que cualesquiera sean $-\infty \leq a < b < \infty$ vale que

$$\mathbb{P}(a < X \leq b) = \int_a^b f_X(x) dx. \quad (11)$$

En particular, para cada $x \in \mathbb{R}$, vale que

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt. \quad (12)$$

Nota Bene. Notar que de (12) se deduce que

$$\int_{-\infty}^{\infty} f_X(x) dx = 1.$$

Aplicando en (12) el teorema Fundamental del Cálculo Integral, se obtiene que si X es *absolutamente continua*, $F_X(x)$ es una función continua para todo x , y su derivada es $f_X(x)$ en todos los x donde f_X es continua.

Como la expresión “absolutamente continua” es demasiado larga, se suele hablar simplemente de “distribuciones continuas”. Sin embargo, hay que tener en cuenta que el hecho de que F_X sea una *función* continua, *no* implica que la distribución de X sea *absolutamente continua*: hay funciones monótonas y continuas, que sin embargo no son la primitiva de ninguna función. (Para más detalles consultar el ejemplo sobre *distribuciones tipo Cantor* que está en Feller Vol II, p.35-36). \square

Interpretación intuitiva de la densidad de probabilidad. Sea X una variable aleatoria absolutamente continua con función densidad $f_X(x)$ continua. Para cada $\epsilon > 0$ pequeño y para $x \in \mathbb{R}$ vale que

$$\mathbb{P}(x - \epsilon/2 < X \leq x + \epsilon/2) = \int_{x-\epsilon/2}^{x+\epsilon/2} f_X(t) dt \approx f_X(x)\epsilon.$$

Dicho en palabras, la probabilidad de que el valor de X se encuentre en un intervalo de longitud ϵ centrado en x es aproximadamente $f_X(x)\epsilon$. \square

Ejemplos

Ejemplo 1.10. El resultado, X , del lanzamiento de un dado equilibrado (ver Ejemplo 1.1) es una variable aleatoria discreta. Esto resulta evidente de observar que el gráfico de la función de distribución de X (ver Figura 1) que tiene la forma de una escalera con saltos de altura $1/6$ en los puntos $1, 2, 3, 4, 5, 6$. Dicho en otras palabras, toda la masa de la variable aleatoria X está concentrada en el conjunto de los átomos de F_X , $\mathbb{A} = \{1, 2, 3, 4, 5, 6\}$. \square

Ejemplo 1.11 (Números al azar). El resultado de “sortear” un número al azar sobre el intervalo $(0, 1)$ es una variable aleatoria absolutamente continua. La probabilidad del evento $U \leq u$ es igual a la longitud del intervalo $(-\infty, u] \cap (0, 1)$.

Notar que cuando $u \leq 0$ el intervalo $(-\infty, u] \cap (0, 1)$ se reduce al conjunto vacío que por definición tiene longitud 0. Por otra parte, para cualquier $u \in (0, 1)$ se tiene que $(-\infty, u] \cap (0, 1) = (0, u)$ y en consecuencia $\mathbb{P}(U \leq u) = u$; mientras que si $u \geq 1$, $(-\infty, u] \cap (0, 1) = (0, 1)$ de donde sigue que $\mathbb{P}(U \leq u) = 1$. Por lo tanto, la función de distribución de U es

$$F_U(u) = u\mathbf{1}\{0 \leq u < 1\} + \mathbf{1}\{u \geq 1\}.$$

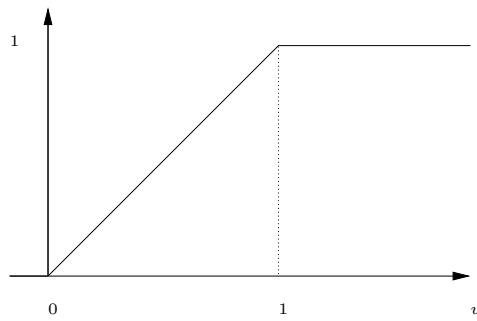


Figura 3: Gráfico de la función de distribución del resultado de “sortear” un número al azar.

Derivando, respecto de u , la función de distribución $F_U(u)$ se obtiene una función densidad para U :

$$f_U(u) = \mathbf{1}\{0 < u < 1\}.$$

□

Nota Bene. Sortear un número al azar sobre el intervalo $(0, 1)$ es un caso particular de una familia de variables aleatorias denominadas *uniformes*. Una variable aleatoria X , definida sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, se denomina *uniformemente distribuida sobre el intervalo* (a, b) , donde $a < b$, si X es absolutamente continua y admite una función densidad de la forma

$$f_X(x) = \frac{1}{b-a}\mathbf{1}\{x \in (a, b)\}.$$

En tal caso escribiremos $X \sim \mathcal{U}(a, b)$.

□

Comentario. En la Sección 1.4 mostraremos que todas las variables aleatorias se pueden construir utilizando variables aleatorias uniformemente distribuidas sobre el intervalo $(0, 1)$.

Ejemplo 1.12. El tiempo, T , de funcionamiento hasta la aparición de la primera falla para un sistema con función intensidad de fallas continua $\lambda(t)$ (ver Ejemplo 1.2) es una variable aleatoria absolutamente continua que admite una densidad de la forma

$$f_T(t) = \lambda(t) \exp\left(-\int_0^t \lambda(s)ds\right) \mathbf{1}\{t > 0\}. \quad (13)$$

Nota Bene: algunos casos particulares del Ejemplo 1.12. El comportamiento de la densidad (13) depende de la forma particular de la función intensidad de fallas $\lambda(t)$. En lo que sigue mostraremos algunos casos particulares.

- *Exponencial de intensidad λ .* Se obtiene poniendo $\lambda(t) = \lambda \mathbf{1}\{t \geq 0\}$, donde λ es una constante positiva, arbitraria pero fija.

$$f_T(t) = \lambda \exp(-\lambda t) \mathbf{1}\{t > 0\}. \quad (14)$$

- *Weibull de parámetros c y α .* Se obtiene poniendo $\lambda(t) = \frac{c}{\alpha} \left(\frac{t}{\alpha}\right)^{c-1} \mathbf{1}\{t \geq 0\}$, donde $c > 0$ y $\alpha > 0$. En este caso, la densidad (13) adopta la forma

$$f_T(t) = \frac{c}{\alpha} \left(\frac{t}{\alpha}\right)^{c-1} \exp\left(-\left(\frac{t}{\alpha}\right)^c\right). \quad (15)$$

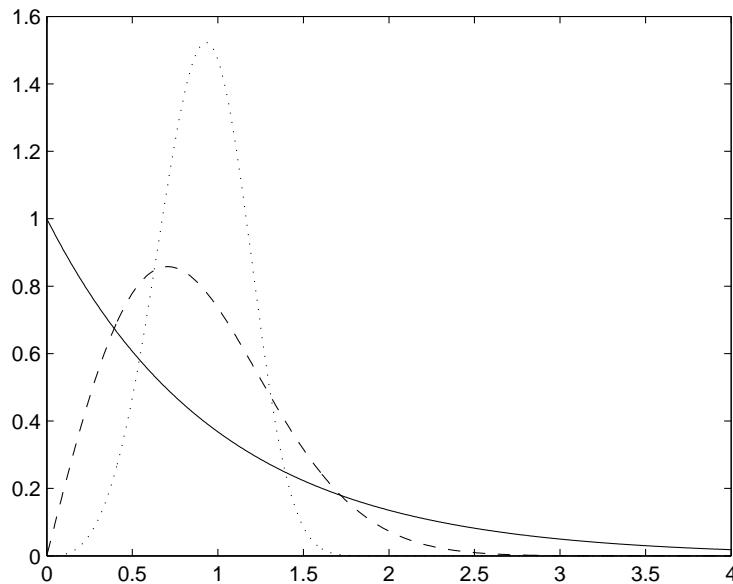


Figura 4: Gráficos de las densidades Weibull de parámetro de escala $\alpha = 1$ y parámetro de forma: $c = 1, 2, 4$: en línea sólida $c = 1$; en línea quebrada $c = 2$ y en línea punteada $c = 4$.

Notar que la exponencial de intensidad λ es un caso especial de la Weibull puesto que (14) se obtiene de (15) poniendo $c = 1$ y $\alpha = \lambda^{-1}$. \square

Ejemplo 1.13. La variable aleatoria, S , considerada en el Ejemplo 1.3 es una variable aleatoria mixta (ver Figura 2) porque no es discreta ni continua. Tiene un único átomo en $s = 0$ y su peso es $\exp\left(-\int_0^{t_0} \lambda(x)dx\right)$. \square

1.3. Cuantiles

Definición 1.14. Sea $\alpha \in (0, 1)$. Un cuantil- α de X es cualquier número $x_\alpha \in \mathbb{R}$ tal que

$$\mathbb{P}(X < x_\alpha) \leq \alpha \quad y \quad \alpha \leq \mathbb{P}(X \leq x_\alpha). \quad (16)$$

Observación 1.15. Notar que las desigualdades que caracterizan a los cuantiles- α se pueden reescribir de la siguiente manera

$$F_X(x_\alpha) - \mathbb{P}(X = x_\alpha) \leq \alpha \quad y \quad \alpha \leq F_X(x_\alpha). \quad (17)$$

Por lo tanto, si $F_X(x)$ es continua, x_α es un cuantil α si y sólo si

$$F_X(x_\alpha) = \alpha. \quad (18)$$

Interpretación “geométrica” del cuantil- α . Si X es una variable aleatoria absolutamente continua con función de densidad $f_X(x)$ el cuantil- α de X es la única solución de la ecuación

$$\int_{-\infty}^{x_\alpha} f_X(x) dx = \alpha.$$

Esto significa que el cuantil- α de X es el único punto sobre el eje de las abscisas a cuya izquierda el área bajo la función de densidad $f_X(x)$ es igual a α .

Nota Bene. Sea $x \in \mathbb{R}$. Las desigualdades (17) significan que x es un cuantil- α si y sólo si $\alpha \in [F(x) - \mathbb{P}(X = x), F(x)]$

Nota Bene. El cuantil- α siempre existe. Sea $\alpha \in (0, 1)$, la existencia del cuantil α se deduce analizando el conjunto $R_X^\alpha = \{x \in \mathbb{R} : \alpha \leq F_X(x)\}$.

1. R_X^α es no vacío porque $\lim_{x \rightarrow \infty} F_X(x) = 1$.
2. R_X^α es acotado inferiormente porque $\lim_{x \rightarrow -\infty} F_X(x) = 0$.
3. Si $x_0 \in R_X^\alpha$, entonces $[x_0, +\infty) \subset R_X^\alpha$ porque $F_X(x)$ es no decreciente.
4. $\inf R_X^\alpha \in R_X^\alpha$ porque existe una sucesión $\{x_n : n \in \mathbb{N}\} \subset R_X^\alpha$ tal que $x_n \downarrow \inf R_X^\alpha$ y $F_X(x)$ es una función continua a derecha:

$$\alpha \leq \lim_{n \rightarrow \infty} F_X(x_n) = F_X\left(\lim_{n \rightarrow \infty} x_n\right) = F_X(\inf R_X^\alpha).$$

De las propiedades anteriores se deduce que

$$R_X^\alpha = [\inf R_X^\alpha, +\infty) = [\min R_X^\alpha, +\infty).$$

Hay dos casos posibles: (a) $F_X(\min R_X^\alpha) = \alpha$ o (b) $F_X(\min R_X^\alpha) > \alpha$.

- (a) Si $F_X(\min R_X^\alpha) = \alpha$, entonces $\mathbb{P}(X < \min R_X^\alpha) = \alpha - \mathbb{P}(X = \min R_X^\alpha) \leq \alpha$.

(b) Si $F_X(\min R_X^\alpha) > \alpha$, entonces

$$\mathbb{P}(X < x) < \alpha \quad \forall x < \min R_X^\alpha \quad (19)$$

porque sino existe un $x < \min R_X^\alpha$ tal que $\alpha \leq \mathbb{P}(X < x) \leq F_X(x)$ y por lo tanto, $x \in R_X^\alpha$ lo que constituye un absurdo.

De (19) se deduce que $\mathbb{P}(X < \min R_X^\alpha) = \lim_{x \uparrow \min R_X^\alpha} F_X(x) \leq \alpha$.

En cualquiera de los dos casos

$$x_\alpha = \min \{x \in \mathbb{R} : F_X(x) \geq \alpha\} \quad (20)$$

es un cuantil- α . □

Nota Bene. Si F_X es discontinua, (18) no tiene siempre solución; y por eso es mejor tomar (16) como definición. Si F_X es estrictamente creciente, los cuantiles son únicos. Pero si no, los valores que satisfacen (18) forman un intervalo.

Cuartiles y mediana. Los cuantiles correspondientes a $\alpha = 0.25, 0.50$ y 0.75 son respectivamente el primer, el segundo y tercer *cuartil*. El segundo cuartil es la *mediana*.

Ejemplos

Ejemplo 1.16. En el Ejemplo 1.1 hemos visto que la función de distribución del resultado del lanzamiento de un dado equilibrado es una escalera con saltos de altura $1/6$ en los puntos $1, 2, 3, 4, 5, 6$:

$$F_X(x) = \sum_{i=1}^5 \frac{i}{6} \mathbf{1}\{i \leq x < i+1\} + \mathbf{1}\{6 \leq x\}.$$

Como la imagen de F_X es el conjunto $\{0, 1/6, 2/6, 3/6, 4/6, 5/6, 1\}$ la ecuación (18) solo tiene solución para $\alpha \in \{1/6, 2/6, 3/6, 4/6, 5/6\}$. Más aún, para cada $i = 1, \dots, 5$

$$F_X(x) = \frac{i}{6} \iff x \in [i, i+1).$$

En otras palabras, para cada $i = 1, \dots, 5$ los cuantiles- $i/6$ de X son el intervalo $[i, i+1)$. En particular, “la” mediana de X es cualquier punto del intervalo $[3, 4]$.

Para cada $\alpha \in (\frac{i-1}{6}, \frac{i}{6})$, $i = 1, \dots, 6$, el cuantil α de X es $x_\alpha = i$. □

Ejemplo 1.17. Sea T el tiempo de funcionamiento hasta la aparición de la primera falla para un sistema con función intensidad de fallas $\lambda(t) = 2t\mathbf{1}\{t \geq 0\}$ (ver Ejemplo 1.2). La función de distribución de T es

$$F_T(t) = \left(1 - \exp\left(-\int_0^t 2sds\right)\right) \mathbf{1}\{t > 0\} = (1 - \exp(-t^2)) \mathbf{1}\{t > 0\}. \quad (21)$$

Como $F_T(t)$ es continua los cuantiles- α , $\alpha \in (0, 1)$, se obtienen resolviendo la ecuación (18):

$$F_T(t) = \alpha \iff 1 - \exp(-t^2) = \alpha \iff t = \sqrt{-\log(1 - \alpha)}.$$

Por lo tanto, para cada $\alpha \in (0, 1)$ el cuantil- α de T es

$$t_\alpha = \sqrt{-\log(1 - \alpha)}. \quad (22)$$

En particular, la mediana de T es $t_{0.5} = \sqrt{-\log(1 - 0.5)} \approx 0.8325$. □

Ejemplo 1.18. Se considera un sistema con función intensidad de fallas $\lambda(t) = 2t\mathbf{1}\{t \geq 0\}$. El sistema debe prestar servicios durante 1 hora. Si durante ese período el sistema falla, se lo repara y se lo vuelve a utilizar hasta que cumpla con el plazo estipulado. Sea S el tiempo de funcionamiento (medido en horas) del sistema después de la primera reparación.

En el Ejemplo 1.3 vimos que la función de distribución de S es

$$\begin{aligned} F_S(s) &= \exp\left(-\int_0^{1-s} 2tdt\right) \mathbf{1}\{0 \leq s < 1\} + \mathbf{1}\{s \geq 1\} \\ &= \exp(-(1-s)^2) \mathbf{1}\{0 \leq s < 1\} + \mathbf{1}\{s \geq 1\}, \end{aligned}$$

y que S es una variable aleatoria mixta (ver Figura 2) con un único átomo en $s = 0$ cuyo peso es e^{-1} . En consecuencia, $s = 0$ es un cuantil- α de S para todo $\alpha \in (0, e^{-1}]$. Restringida al intervalo $(0, 1)$ la función $F_S(s)$ es continua y su imagen es el intervalo $(e^{-1}, 1)$. Por ende, para cada $\alpha \in (e^{-1}, 1)$ el cuantil- α de S se obtiene resolviendo la ecuación $F_S(s) = \alpha$:

$$\begin{aligned} F_S(s) = \alpha &\iff \exp(-(1-s)^2) = \alpha \iff -(1-s)^2 = \log(\alpha) \\ &\iff (1-s)^2 = -\log(\alpha) \iff |1-s| = \sqrt{-\log(\alpha)} \\ &\iff 1-s = \sqrt{-\log(\alpha)} \iff 1 - \sqrt{-\log(\alpha)} = s. \end{aligned}$$

Por lo tanto, para cada $\alpha \in (e^{-1}, 1)$ el cuantil- α de S es

$$s_\alpha = 1 - \sqrt{-\log(\alpha)}.$$

En particular, la mediana de S es $s_{0.5} = 1 - \sqrt{-\log(0.5)} \approx 0.1674$. \square

1.4. Construcción de variables aleatorias

Teorema 1.19 (Simulación). Sea $F : \mathbb{R} \rightarrow [0, 1]$ una función con las siguientes propiedades

(F1) es *no decreciente*: si $x_1 \leq x_2$, entonces $F(x_1) \leq F(x_2)$;

(F2) es *continua a derecha*: para todo $x_0 \in \mathbb{R}$ vale que $\lim_{x \downarrow x_0} F(x) = F(x_0)$;

(F3) $\lim_{x \rightarrow -\infty} F(x) = 0$ y $\lim_{x \rightarrow \infty} F(x) = 1$.

Existe una variable aleatoria X tal que $F(x) = \mathbb{P}(X \leq x)$.

Esquema de la demostración.

1º) Definir la *inversa generalizada* de F mediante

$$F^{-1}(u) := \min\{x \in \mathbb{R} : u \leq F(x)\}, \quad u \in (0, 1).$$

2º) Definir X mediante

$$X := F^{-1}(U), \quad \text{donde } U \sim \mathcal{U}(0, 1).$$

3º) Observar que vale la equivalencia (inmediata) $F^{-1}(u) \leq x \Leftrightarrow u \leq F(x)$ y deducir que $\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$. \square

Observación 1.20. Si la función F del enunciado del Teorema 1.19 es continua, la inversa generalizada es simplemente la inversa. \square

Nota Bene. El esquema de la demostración del Teorema 1.19 muestra *cómo se construye una variable aleatoria X con función de distribución $F_X(x)$* . La construcción es clave para simular variables aleatorias en una computadora: algoritmos estándar generan variables aleatorias U con distribución uniforme sobre el intervalo $(0, 1)$, aplicando la inversa generalizada de la función de distribución se obtiene la variable aleatoria $F_X^{-1}(U)$ cuya función de distribución es $F_X(x)$. \square

Método gráfico para calcular inversas generalizadas. Sea $u \in (0, 1)$, por definición, $F^{-1}(u) := \min\{x \in \mathbb{R} : u \leq F(x)\}$, $0 < u < 1$. Gráficamente esto significa que para calcular $F^{-1}(u)$ hay que determinar el conjunto de todos los puntos del gráfico de $F(x)$ que están sobre o por encima de la recta horizontal de altura u y proyectarlo sobre el eje de las abscisas. El resultado de la proyección es una semi-recta sobre el eje de las abscisas y el valor de la abscisa que la cierra por izquierda es el valor de $F^{-1}(u)$. \square

Ejemplo 1.21 (Moneda cargada). *Se quiere simular el lanzamiento de una moneda “cargada” con probabilidad $p \in (0, 1)$ de salir cara.* El problema se resuelve construyendo una variable aleatoria X a valores $\{0, 1\}$ tal que $\mathbb{P}(X = 1) = p$ y $\mathbb{P}(X = 0) = 1 - p$, ($X = 1$ representa el evento “la moneda sale cara” y $X = 0$ “la moneda sale ceca”). La función de distribución de X debe ser $F(x) = (1 - p)\mathbf{1}\{0 \leq x < 1\} + \mathbf{1}\{x \geq 1\}$ y su gráfico se muestra en la Figura 5.

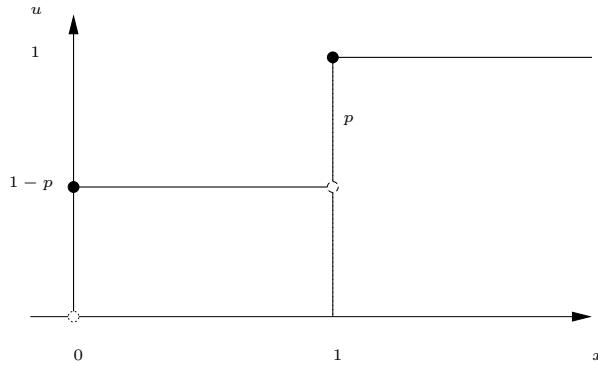


Figura 5: Gráfico de la función $F(x) = (1 - p)\mathbf{1}\{0 \leq x < 1\} + \mathbf{1}\{x \geq 1\}$.

La demostración del Teorema 1.19 indica que para construir la variable aleatoria X lo primero que hay que hacer es determinar la expresión de la inversa generalizada de $F(x)$. Para ello usaremos el método gráfico.

En la Figura 5 se puede ver que para cada $0 < u \leq 1 - p$ el conjunto $\{x \in \mathbb{R} : u \leq F(x)\}$ es la semi-recta $[0, \infty)$ y el punto que la cierra por izquierda es $x = 0$. En consecuencia $F^{-1}(u) = 0$ para todo $0 < u \leq 1 - p$. Del mismo modo se puede ver que $F^{-1}(u) = 1$ para todo $1 - p < u < 1$. Por lo tanto, $F^{-1}(u) = \mathbf{1}\{1 - p < u < 1\}$.

Definiendo $X := \mathbf{1}\{1 - p < U < 1\}$, donde $U \sim \mathcal{U}(0, 1)$ se obtiene la variable aleatoria deseada.

Ejemplo 1.22 (Moneda cargada). *Simular diez lanzamientos de una moneda “cargada” con probabilidad 0.6 de salir cara en cada lanzamiento.*

De acuerdo con el resultado obtenido en el Ejemplo 1.21, para simular el lanzamiento de una moneda cargada con probabilidad 0.6 de salir cara se construye la variable aleatoria $X := \mathbf{1}\{0.4 < U < 1\}$, donde $U \sim \mathcal{U}(0, 1)$.

Para simular 10 valores de X se simulan 10 valores de U . Si en 10 simulaciones de U se obtuviesen los valores 0.578, 0.295, 0.885, 0.726, 0.548, 0.048, 0.474, 0.722, 0.786, 0.598, los valores de la variable X serían 1, 0, 1, 1, 1, 0, 1, 1, 1, respectivamente, y en tal caso, los resultados de los 10 lanzamientos de la moneda serían $H, T, H, H, H, H, T, H, H, H$. \square

Ejemplo 1.23 (Fiabilidad). *Se considera un sistema electrónico con función intensidad de fallas de la forma $\lambda(t) = 2t\mathbf{1}\{t > 0\}$. Se quiere estimar la función de probabilidad de la cantidad de fallas ocurridas durante la primer unidad de tiempo de funcionamiento.*

Para simplificar el problema vamos a suponer que cada vez que se produce una falla, el sistema se repara instantáneamente renovándose sus condiciones iniciales de funcionamiento. Según el Ejemplo 1.2, la función de distribución del tiempo de funcionamiento hasta la aparición de la primer falla es

$$F(t) = (1 - \exp(-t^2)) \mathbf{1}\{t > 0\}. \quad (23)$$

Debido a que la función de distribución $F(t)$ es continua, su inversa generalizada es simplemente su inversa y se obtiene despejando t de la ecuación $1 - \exp(-t^2) = u$. En consecuencia, $F^{-1}(u) = \sqrt{-\log(1-u)}$, $u \in (0, 1)$. Para construir la variable T usamos un número aleatorio U , uniformemente distribuido sobre el intervalo $(0, 1)$ y definimos

$$T := F^{-1}(U) = \sqrt{-\log(1-U)}. \quad (24)$$

La ventaja de la construcción es que puede implementarse casi de inmediato en una computadora. Por ejemplo, una rutina en Octave para simular T es la siguiente

```
U=rand;
T=sqrt(-log(1-rand))
```

Sobre la base de esa rutina podemos simular valores de T . Por ejemplo, en diez simulaciones de T obtuvimos los valores siguientes: 0.3577, 1.7233, 1.1623, 0.3988, 1.4417, 0.3052, 1.1532, 0.3875, 0.8493, 0.9888.

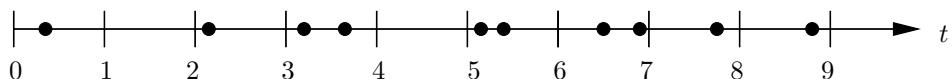


Figura 6: Simulación de los tiempos de ocurrencia de las fallas de un sistema electrónico con función intensidad de fallas de la forma $\lambda(t) = 2t\mathbf{1}\{t \geq 0\}$. Las fallas ocurren los instantes 0.3577, 2.0811, 3.2434, 3.6422, 5.0839, 5.3892, 6.5423, 6.9298, 7.7791, 8.7679.

La rutina puede utilizarse para simular cien mil realizaciones del experimento que consiste en observar la cantidad de fallas durante la primer unidad de tiempo de funcionamiento del sistema electrónico bajo consideración: $N[0, 1] := \min\{n \geq 1 : \sum_{i=1}^n T_i > 1\} - 1$, donde T_1, T_2, \dots son realizaciones independientes de los tiempos de funcionamiento del sistema hasta la ocurrencia de una falla.

Por ejemplo, repitiendo la simulación 100000 veces obtuvimos la siguiente tabla que contiene la cantidad de veces que fué simulado cada valor de la variable $N[0, 1]$:

valor simulado	0	1	2	3	4
frecuencia	36995	51792	10438	743	32

(25)

obteniéndose las siguientes estimaciones

$$\begin{aligned}\mathbb{P}(N[0, 1] = 0) &\approx 0.36995, \quad \mathbb{P}(N[0, 1] = 1) \approx 0.51792, \quad \mathbb{P}(N[0, 1] = 2) \approx 0.10438, \\ \mathbb{P}(N[0, 1] = 3) &\approx 0.00743, \quad \mathbb{P}(N[0, 1] = 4) \approx 0.00032.\end{aligned}$$

Para finalizar este ejemplo, presentamos una rutina en Octave que simula cien mil veces la cantidad de fallas en la primer unidad de tiempo y que al final produce los resultados para construir una tabla similar a la tabla (25).

```
for i=1:100000
    n=-1;
    S=0;
    while S<=1;
        T=sqrt(-log(1-rand));
        S=S+T;
        n=n+1;
    end
    f(i)=n;
end
M=max(f);
for i=1:M+1;
    N(i)=length(find(f==i-1));
end
N
```

Ejemplo 1.24 (*Saltando, saltando, sa, sa, sa, saltando, ... ↴*). La función

$$F(x) = \sum_{n=1}^{\infty} \frac{1}{2^n} \mathbf{1}\{x \geq r_n\}, \quad (26)$$

donde r_1, r_2, \dots es un reordenamiento de los números racionales del intervalo $(0, 1)$ con denominadores crecientes: $\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \dots$, tiene las siguientes propiedades: es creciente, continua a derecha, $\lim_{x \rightarrow -\infty} F(x) = 0$ y $\lim_{x \rightarrow \infty} F(x) = 1$; tiene saltos en todos los números racionales del $(0, 1)$ y es continua en los irracionales del $(0, 1)$.

Pero no! Mejor no hablar de ciertas cosas ...

□

Ejercicios adicionales

3. Sea X una variable aleatoria con función de distribución $F_X(x)$. Mostrar que para cada $\alpha \in (0, 1)$ vale que

$$\sup\{x \in \mathbb{R} : F_X(x) < \alpha\} = \min\{x \in \mathbb{R} : F_X(x) \geq \alpha\}.$$

1.5. Función de distribución empírica e histogramas

Distribución empírica

La función de distribución empírica $F_n(x)$ de n puntos sobre la recta x_1, \dots, x_n es la función escalera con saltos de altura $1/n$ en los puntos x_1, \dots, x_n . En otras palabras, $nF_n(x)$ es igual a la cantidad de puntos x_k en $(-\infty, x]$ y $F_n(x)$ es una función de distribución:

$$F_n(x) = \frac{1}{n} |\{i = 1, \dots, n : x_i \leq x\}| = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i \leq x\}. \quad (27)$$

Nota Bene. En la práctica, disponemos de conjuntos de observaciones (“muestras”) correspondientes a un experimento considerado aleatorio y queremos extraer de ellas conclusiones sobre los modelos que podrían cumplir. Dada una muestra x_1, \dots, x_n , la función de distribución empírica $F_n(x)$ coincide con la función de distribución de una variable aleatoria discreta que concentra toda la masa en los valores x_1, \dots, x_n , dando a cada uno probabilidad $1/n$.

Observación 1.25. Sea $F_n(x)$ la función de distribución empírica correspondiente a una muestra de n valores x_1, \dots, x_n . Sean a y b dos números reales tales que $a < b$. Notar que

$$F_n(b) - F_n(a) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i \in (a, b]\} = \frac{1}{n} |\{i = 1, \dots, n : x_i \in (a, b]\}|.$$

En consecuencia, el cociente incremental de $F_n(x)$ sobre el intervalo $[a, b]$ es la frecuencia relativa de los valores de la muestra x_1, \dots, x_n contenidos en el intervalo $(a, b]$ “normalizada” por la longitud de dicho intervalo:

$$\frac{F_n(b) - F_n(a)}{b - a} = \left(\frac{1}{b - a} \right) \left(\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i \in (a, b]\} \right). \quad (28)$$

Notar que si los n valores, x_1, \dots, x_n , corresponden a n observaciones independientes de los valores de una variable aleatoria X , la interpretación intuitiva de la probabilidad indica que el cociente incremental (28) debería estar próximo del cociente incremental de la función de distribución, $F_X(x)$, de la variable aleatoria X sobre el intervalo $[a, b]$:

$$\frac{F_n(b) - F_n(a)}{b - a} \approx \frac{\mathbb{P}(a < X \leq b)}{b - a} = \frac{F_X(b) - F_X(a)}{b - a}. \quad (29)$$

Cuando X es una variable aleatoria absolutamente continua con función densidad continua $f_X(x)$ la aproximación (28) adopta la forma

$$\frac{F_n(b) - F_n(a)}{b - a} \approx \frac{1}{b - a} \int_a^b f_X(x) dx = f_X(x), \quad (30)$$

donde x es algún punto perteneciente al intervalo (a, b) . □

Histogramas

Un *histograma* de una muestra x_1, \dots, x_n se obtiene eligiendo una partición en m intervalos de extremos $a_0 < \dots < a_m$, con longitudes $L_j = a_j - a_{j-1}$; calculando las *frecuencias relativas*

$$p_j = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{a_{j-1} < x_i < a_j\}$$

y graficando la función igual a p_j/L_j en el intervalo $(a_{j-1}, a_j]$ y a 0 fuera de los intervalos:

$$f_{x_1, \dots, x_n; a_0, \dots, a_m}(x) := \sum_{j=1}^m \frac{p_j}{L_j} \mathbf{1}\{x \in (a_{j-1}, a_j]\}. \quad (31)$$

O sea, un conjunto de rectángulos con área p_j .

Cuando la muestra x_1, \dots, x_n corresponde a n observaciones independientes de una variable aleatoria X absolutamente continua la función definida en (31) es una versión discreta de la densidad de X en la que las áreas miden frecuencias relativas.

Ejercicios adicionales

4. Lucas filma videos de tamaños aleatorios. En una muestra aleatoria de 5 videos filmados por Lucas se obtuvieron los siguiente tamaños (en MB):

$$17, 21.3, 18.7, 21, 18.7$$

Hallar y graficar la función de distribución empírica asociada a esta muestra. Estimar, usando la función de distribución empírica asociada a esta muestra, la probabilidad de que un video ocupe menos de 19.5 MB.

5. Los siguientes datos corresponden a los tiempos de funcionamiento (en años) hasta que ocurre la primer falla de una muestra de 12 máquinas industriales:

$$\begin{aligned} & 2.0087, 1.9067, 2.0195, 1.9242, 1.8885, 1.8098, \\ & 1.9611, 2.0404, 2.1133, 2.0844, 2.1695, 1.9695. \end{aligned}$$

Usando los intervalos con extremos 1.7, 1.9, 2.1, 2.3, hallar la función histograma basada en la muestra observada e integrarla para estimar la probabilidad de que una máquina industrial del mismo tipo funcione sin fallas durante menos de dos años.

Ejemplo 1.26. Sea T una variable aleatoria con distribución exponencial de intensidad 1 (ver (14)). Esto es, T es una variable aleatoria absolutamente continua con función densidad de probabilidad

$$f_T(t) = e^{-t} \mathbf{1}\{t > 0\}$$

y función de distribución

$$F_T(t) = (1 - e^{-t}) \mathbf{1}\{t \geq 0\}.$$

De acuerdo con el esquema de la demostración del Teorema 1.19 podemos simular muestras de T utilizando un generador de números aleatorios uniformemente distribuidos sobre el intervalo $(0, 1)$. Concretamente, si $U \sim \mathcal{U}(0, 1)$, entonces

$$\hat{T} = -\log(1 - U)$$

es una variable con distribución exponencial de intensidad 1.

Para obtener una muestra de 10 valores t_1, \dots, t_{10} de una variable con distribución exponencial de intensidad 1 generamos 10 números aleatorios u_1, \dots, u_{10} y los transformamos poniendo $t_i = -\log(1 - u_i)$. Por ejemplo, si los valores u_1, \dots, u_{10} son, respectivamente,

$$0.1406, 0.3159, 0.8613, 0.4334, 0.0595, 0.8859, 0.2560, 0.2876, 0.2239, 0.5912,$$

los valores de la muestra obtenida, t_1, \dots, t_{10} , son, respectivamente,

$$0.1515, 0.3797, 1.9753, 0.5682, 0.0613, 2.1703, 0.2957, 0.3390, 0.2535, 0.8946. \quad (32)$$

La función de distribución empírica de la muestra observada, $F_{10}(t)$, es una función escalera con saltos de altura $1/10$ en los siguientes puntos del eje t :

$$0.0613, 0.1515, 0.2535, 0.2957, 0.3390, 0.3797, 0.5682, 0.8946, 1.9753, 2.1703.$$

Para construir un histograma usaremos la partición que se obtiene dividiendo en dos intervalos de igual longitud el intervalo comprendido entre los valores mínimos y máximos observados: 0.0613, 1.1158, 2.1703. La longitud L de cada intervalo es 1.0545. La frecuencia relativa de la muestra sobre el primer intervalo es $p_1 = 8/10$ y sobre el segundo $p_2 = 2/10$ y la correspondiente altura de cada rectángulo es $p_1/L = 0.75865$ y $p_2/L = 0.18966$.

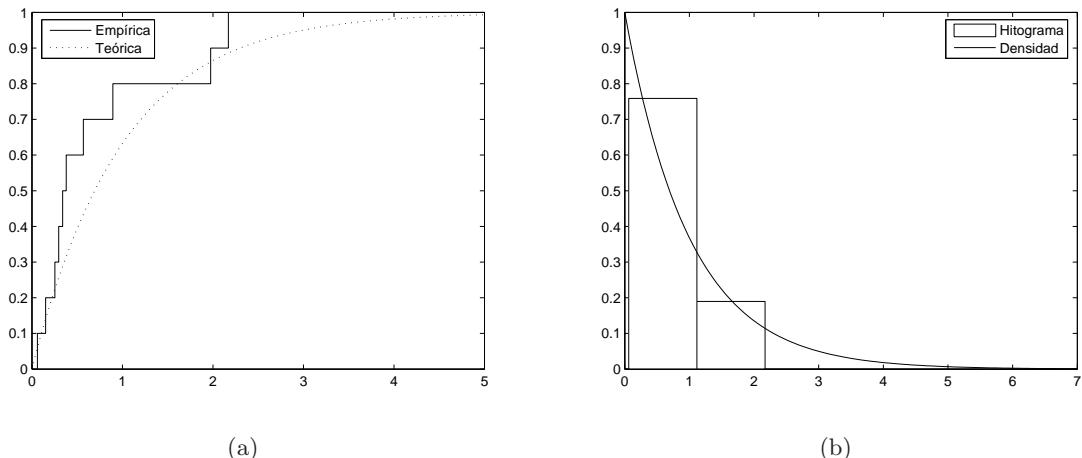


Figura 7: (a) Gráficos de la función de distribución empírica $F_{10}(t)$ correspondiente a la muestra dada en (32) y de la función de distribución de T . (b) Histograma correspondiente a la misma muestra y gráfico de la densidad de T .

Para producir los gráficos de la Figura 7 usamos las siguientes rutinas en Octave.

Rutina para simular 10 valores de una exponencial de intensidad 1

```
U=rand(1,10);
T=-log(1-U);
```

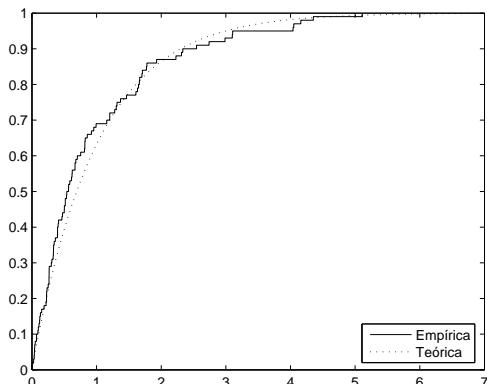
Rutina para graficar la función de distribución empírica de la muestra T

```
t=sort(T);
s=empirical_cdf(t,t);
stairs([t(1),t],[0 s])
```

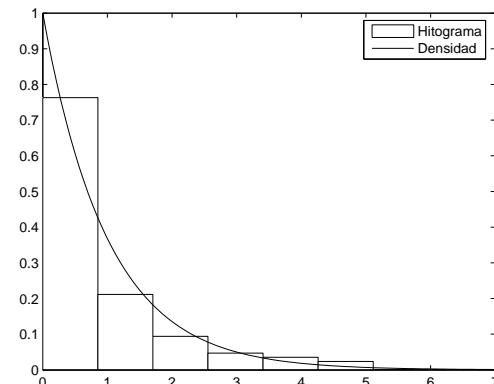
Rutina para graficar un histograma de la muestra T

```
[f,c]=hist(T,2);
p=f/10;
L=c(2)-c(1);
bar(c,p/L,1,'w')
```

Usando rutinas similares para muestras de tamaño 100 se obtienen los siguientes gráficos.



(a)



(b)

Figura 8: (a) Gráficos de la función de distribución empírica $F_{100}(t)$ correspondiente a una muestra de tamaño 100 de una variable T con distribución exponencial de intensidad 1 y de la función de distribución de T . (b) Histograma correspondiente a la misma muestra y gráfico de la densidad de T . \square

2. Variables truncadas

Sea X una variable aleatoria definida sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Sea $B \subset \mathbb{R}$ un conjunto tal que $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}$ y tal que $\mathbb{P}(X \in B) > 0$.

Truncar la variable aleatoria X al conjunto B significa condicionarla a tomar valores en el conjunto B .

Mediante $X|X \in B$ designaremos la variable aleatoria obtenida por truncar X al conjunto B . Por definición, la función de distribución de $X|X \in B$ es

$$F_{X|X \in B}(x) = \mathbb{P}(X \leq x | X \in B) = \frac{\mathbb{P}(X \leq x, X \in B)}{\mathbb{P}(X \in B)}. \quad (33)$$

Caso absolutamente continuo. Si la variable aleatoria X es absolutamente continua con densidad de probabilidades $f_X(x)$, la función de distribución de $X|X \in B$ adopta la forma

$$F_{X|X \in B}(x) = \frac{\int_{\{X \leq x\} \cap \{X \in B\}} f_X(x) dx}{\mathbb{P}(X \in B)} = \frac{\int_{-\infty}^x f_X(x) \mathbf{1}\{x \in B\} dx}{\mathbb{P}(X \in B)}. \quad (34)$$

Por lo tanto, $X|X \in B$ es una variable aleatoria absolutamente continua con densidad de probabilidades

$$f_{X|X \in B}(x) = \frac{f_X(x)}{\mathbb{P}(X \in B)} \mathbf{1}\{x \in B\}. \quad (35)$$

Nota Bene. La densidad condicional $f_{X|X \in B}(x)$ es cero fuera del conjunto condicionante B . Dentro del conjunto condicionante la densidad condicional tiene exactamente la misma forma que la densidad incondicional, salvo que está escalada por el factor de normalización $1/\mathbb{P}(X \in B)$ que asegura que $f_{X|X \in B}(x)$ integra 1. \square

Ejemplo 2.1 (Exponencial truncada a la derecha). Sea T una variable aleatoria con distribución exponencial de intensidad $\lambda > 0$ y sea $t_0 > 0$. Según la fórmula (35) la variable aleatoria T truncada a la semi-recta $(t, +\infty)$, $T|T > t_0$, tiene la siguiente densidad de probabilidades

$$f_{T|T>t_0}(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t_0}} \mathbf{1}\{t > t_0\} = e^{-\lambda(t-t_0)} \mathbf{1}\{t - t_0 > 0\} = f_T(t - t_0).$$

En otros términos, si $T \sim \text{Exp}(\lambda)$, entonces $T|T > t_0 \sim t_0 + \text{Exp}(\lambda)$. \square

Caso discreto. El caso discreto se trata en forma análoga a la anterior. La función de probabilidad de $X|X \in B$ adopta la forma

$$p_{X|X \in B}(x) = \frac{\mathbb{P}(X = x)}{\mathbb{P}(X \in B)} \mathbf{1}\{x \in B\}. \quad (36)$$

Ejemplo 2.2 (Dado equilibrado). Sea X el resultado del tiro de un dado equilibrado y sea $B = \{2, 4, 6\}$. El evento “el resultado del tiro es un número par” es $X \in B$. Aplicando la fórmula anterior obtenemos

$$p_{X|X \in B}(x) = \frac{1/6}{1/2} \mathbf{1}\{x \in \{2, 4, 6\}\} = \frac{1}{3} \mathbf{1}\{x \in \{2, 4, 6\}\}. \quad (37)$$

\square

2.1. Perdida de memoria

Ejemplo 2.3. Lucas camina hacia la parada del colectivo. El tiempo, T , entre llegadas de colectivos tiene distribución exponencial de intensidad λ . Supongamos que Lucas llega t minutos después de la llegada de un colectivo. Sea X el tiempo que Lucas tendrá que esperar hasta que llegue el próximo colectivo. Cuál es la distribución del tiempo de espera X ?

Designamos mediante $A = \{T > t\}$ el evento “*Lucas llegó t minutos después de la llegada de un colectivo*”. Tenemos que

$$\begin{aligned}\mathbb{P}(X > x | A) &= \mathbb{P}(T > t + x | T > t) = \frac{\mathbb{P}(T > t + x, T > t)}{\mathbb{P}(T > t)} \\ &= \frac{\mathbb{P}(T > t + x)}{\mathbb{P}(T > t)} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x}.\end{aligned}$$

□

Definición 2.4. Se dice que una variable aleatoria T no tiene memoria, o *pierde memoria*, si

$$\mathbb{P}(T > s + t | T > t) = \mathbb{P}(T > s) \quad \text{para todo } s, t \geq 0. \quad (38)$$

La condición de pérdida de memoria es equivalente a la siguiente

$$\mathbb{P}(T > s + t) = \mathbb{P}(T > s)\mathbb{P}(T > t). \quad (39)$$

En efecto, basta observar que $\mathbb{P}(T > s + t, T > t) = \mathbb{P}(T > s + t)$ y usar la definición de probabilidad condicional.

Nota Bene. Si se piensa que T es el tiempo para completar cierta operación, la ecuación (38) establece que si a tiempo t la operación no ha sido completada, la probabilidad de que la operación no se complete a tiempo $s + t$ es la misma que la probabilidad inicial de que la operación no haya sido completada a tiempo s . □

Lema 2.5. *La variable exponencial no tiene memoria.*

Demostración Si $T \sim \text{Exp}(\lambda)$, entonces

$$\mathbb{P}(T > t) = e^{-\lambda t} \quad \text{para todo } t \geq 0. \quad (40)$$

Usando (40) se prueba inmediatamente que la ecuación (39) se satisface cuando T tiene distribución exponencial (pues $e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t}$). □

Nota Bene. Si modelamos el tiempo para completar cierta operación por una variable aleatoria T con distribución exponencial, la propiedad de pérdida de memoria implica que mientras la operación no haya sido completada, el tiempo restante para completarla tiene la misma función de distribución, no importa cuando haya empezado la operación. □

Ejemplo 2.6. Supongamos que el tiempo de espera para recibir un mensaje tenga distribución exponencial de intensidad $1/10$ minutos. Cuál es la probabilidad de que tengamos que esperar más de 15 minutos para recibirllo? Cuál es la probabilidad de que tengamos que esperar más de 15 minutos para recibir el mensaje dado que hace más de 10 minutos que lo estamos esperando?

Si T representa el tiempo de espera, $T \sim \text{Exp}(1/10)$. La primer probabilidad es

$$\mathbb{P}(T > 15) = e^{-\frac{1}{10} \cdot 15} = e^{-\frac{3}{2}} \approx 0.220$$

La segunda pregunta interroga por la probabilidad de que habiendo esperado 10 minutos tengamos que esperar al menos 5 minutos más. Usando la propiedad de falta de memoria de la exponencial, dicha probabilidad es

$$\mathbb{P}(T > 5) = e^{-\frac{1}{10} \cdot 5} = e^{-\frac{1}{2}} \approx 0.604.$$

□

2.2. Caracterización cualitativa de la distribución exponencial

La propiedad de pérdida de memoria caracteriza a la distribución exponencial.

Teorema 2.7. Sea T una variable aleatoria continua a valores en \mathbb{R}^+ . Si T pierde memoria, entonces $T \sim \text{Exp}(\lambda)$, donde $\lambda = -\log \mathbb{P}(T > 1)$.

Demostración (a la Cauchy). Sea $G(t) := \mathbb{P}(T > t)$. De la ecuación (39) se deduce que

$$G(s+t) = G(s)G(t). \quad (41)$$

La única función continua a derecha que satisface la ecuación funcional (41) es

$$G(t) = G(1)^t. \quad (42)$$

Para ello basta ver que $G\left(\frac{m}{n}\right) = G(1)^{\frac{m}{n}}$. Si vale (41), entonces $G\left(\frac{2}{n}\right) = G\left(\frac{1}{n} + \frac{1}{n}\right) = G\left(\frac{1}{n}\right)G\left(\frac{1}{n}\right) = G\left(\frac{1}{n}\right)^2$ y repitiendo el argumento se puede ver que

$$G\left(\frac{m}{n}\right) = G\left(\frac{1}{n}\right)^m. \quad (43)$$

En particular, si $m = n$ se obtiene $G(1) = G\left(\frac{1}{n}\right)^n$. Equivalentemente,

$$G\left(\frac{1}{n}\right) = G(1)^{\frac{1}{n}} \quad (44)$$

De las identidades (43) y (44) se deduce que

$$G\left(\frac{m}{n}\right) = G(1)^{\frac{m}{n}}. \quad (45)$$

Ahora bien, debido a que $G(1) = \mathbb{P}(T > 1) \in (0, 1)$, existe $\lambda > 0$ tal que $G(1) = e^{-\lambda}$ ($\lambda = -\log G(1)$). Reemplazando en (42) se obtiene $G(t) = (e^{-\lambda})^t = e^{-\lambda t}$. □

2.3. Dividir y conquistar

Teorema 2.8. Sea X una variable aleatoria absolutamente continua con densidad de probabilidades $f_X(x)$. Sea $(B_i)_{i \geq 1}$ una familia de subconjuntos disjuntos dos a dos de la recta real tales que $\{X \in B_i\} \in \mathcal{A}$ y $\mathbb{P}(X \in B_i) > 0$ para todo $i \geq 1$. Si $\Omega = \cup_{i \geq 1} \{X \in B_i\}$, entonces

$$f_X(x) = \sum_{i \geq 1} f_{X|X \in B_i}(x) \mathbb{P}(X \in B_i). \quad (46)$$

Demostración. Inmediata de la fórmula (35) y de observar que $\sum_{i \geq 1} \mathbf{1}\{X \in B_i\} = 1$. \square

Ejemplo 2.9 (Dividir y conquistar). Todas las mañanas Lucas llega a la estación del subte entre las 7:10 y las 7:30 (con distribución uniforme en el intervalo). El subte llega a la estación cada quince minutos comenzando a las 6:00. ¿Cuál es la densidad de probabilidades del tiempo que tiene que esperar Lucas hasta subirse al subte?

Sea X el tiempo de llegada de Lucas a la estación del subte, $X \sim \mathcal{U}[7:10, 7:30]$. Sea Y el tiempo de espera. Consideramos los eventos $A = \{7:10 \leq X \leq 7:15\} = "Lucas sube en el subte de las 7:15"$; $B = \{7:15 < X \leq 7:30\} = "Lucas sube en el subte de las 7:30"$.

Condicionado al evento A , el tiempo de llegada de Lucas a la estación del subte es uniforme entre las 7:10 y las 7:15. En ese caso, el tiempo de espera Y es uniforme entre 0 y 5 minutos. Análogamente, condicionado al evento B , Y es uniforme entre 0 y 15 minutos. La densidad de probabilidades de Y se obtiene dividiendo y conquistando

$$\begin{aligned} f_Y(y) &= \left(\frac{5}{20}\right) \frac{1}{5} \mathbf{1}\{0 \leq y \leq 5\} + \left(\frac{15}{20}\right) \frac{1}{15} \mathbf{1}\{0 \leq y \leq 15\} \\ &= \frac{1}{10} \mathbf{1}\{0 \leq y \leq 5\} + \frac{1}{20} \mathbf{1}\{5 \leq y \leq 15\}. \end{aligned}$$

\square

3. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Bertsekas, D. P., Tsitsiklis, J. N.: Introduction to Probability. M.I.T. Lecture Notes. (2000)
2. Chung, K. L.: A Course in Probability Theory. Academic Press, San Diego. (2001)
3. Durrett R.: Probability. Theory and Examples. Duxbury Press, Belmont. (1996)
4. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons, New York. (1968)
5. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 2. John Wiley & Sons, New York. (1971)
6. Grimmett, G. R., Stirzaker, D. R.: Probability and Random Processes. Oxford University Press, New York. (2001)
7. Johnson, N. L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions. Vol. 1. John Wiley & Sons, New York. (1995)
8. Kolmogorov, A. N.: Foundations of the Theory of Probability. Chelsea Publishing Co., New York. (1956)
9. Maronna R.: Probabilidad y Estadística Elementales para Estudiantes de Ciencias. Editorial Exacta, La Plata. (1995).
10. Pugachev, V. S.: Introducción a la Teoría de las Probabilidades. Mir, Moscú. (1973)
11. Ross, S.: Introduction to Probability Models. Academic Press, San Diego. (2007)

Vectores aleatorios: marginales e independencia (Borradores, Curso 23)

Sebastian Grynberg

25 de marzo 2013



Um coup de dés jamais n'abolira le hasard
(Stéphane Mallarmé)

Índice

1. Vectores aleatorios	2
1.1. Distribución conjunta	2
1.2. Distribuciones marginales	5
1.2.1. Marginales discretas	5
1.2.2. Marginales continuas	6
1.3. Independencia	8
1.3.1. Caso bidimensional discreto	9
1.3.2. Caso bidimensional continuo	11
2. Bibliografía consultada	12

1. Vectores aleatorios

Notación. Para simplificar la escritura usaremos las siguientes notaciones. Los puntos del espacio n -dimensional \mathbb{R}^n , $n \geq 2$, se denotan en negrita, $\mathbf{x} = (x_1, \dots, x_n)$. La desigualdad $\mathbf{y} \leq \mathbf{x}$ significa que $y_i \leq x_i$ para todo $i = 1, \dots, n$ y se puede interpretar diciendo que \mathbf{y} está al “sudoeste” de \mathbf{x} . El conjunto de todos los puntos al “sudoeste” de \mathbf{x} será denotado mediante $S_{\mathbf{x}} := \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} \leq \mathbf{x}\}$. Finalmente, cualquiera sea el subconjunto de índices $J = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$ denotaremos mediante $\mathbf{x}_J \in \mathbb{R}^m$ al punto m -dimensional que se obtiene de \mathbf{x} quitándole todas las coordenadas que tengan índices fuera de J . Por ejemplo, si $J = \{1, 2\}$, entonces $\mathbf{x}_J = (x_1, x_2)$.

Definición 1.1. Un *vector aleatorio* sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ es una función $\mathbf{X} = (X_1, \dots, X_n) : \Omega \rightarrow \mathbb{R}^n$ tal que para todo $\mathbf{x} \in \mathbb{R}^n$

$$\{\mathbf{X} \in S_{\mathbf{x}}\} = \{\omega \in \Omega : \mathbf{X}(\omega) \leq \mathbf{x}\} \in \mathcal{A}.$$

1.1. Distribución conjunta

La función de distribución (*conjunta*) $F_{\mathbf{X}} : \mathbb{R}^n \rightarrow [0, 1]$ del vector aleatorio \mathbf{X} se define por

$$F_{\mathbf{X}}(\mathbf{x}) := \mathbb{P}(\mathbf{X} \in S_{\mathbf{x}}) \tag{1}$$

Cálculo de probabilidades. La función de distribución conjunta resume toda la información relevante sobre el comportamiento de las variables aleatorias X_1, \dots, X_n . Para fijar ideas, consideremos el caso más simple: $n = 2$. Si $a_1 < b_1$ y $a_2 < b_2$ vale que¹

$$\mathbb{P}(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2). \tag{2}$$

La identidad (2) permite calcular la probabilidad de observar al vector (X_1, X_2) en el rectángulo $(a_1, b_1] \times (a_2, b_2]$.

La fórmula n -dimensional análoga de (2) es complicada y no es relevante para el desarrollo posterior. (Se obtiene aplicando la fórmula de inclusión-exclusión para calcular la probabilidad de la unión de eventos.)

¹Ver la Figura 1.

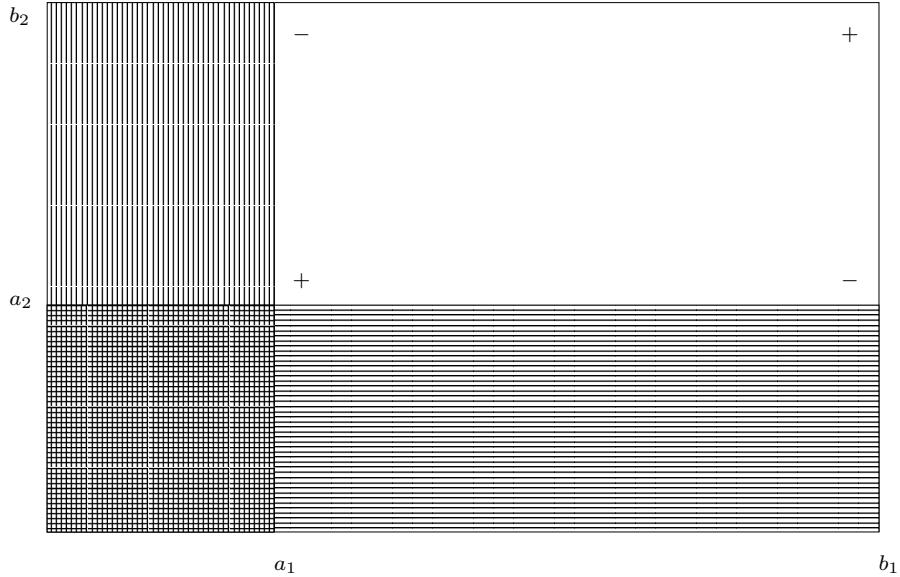


Figura 1: Esquema de la demostración de la identidad (2). El rectángulo $(a_1, b_1] \times (a_2, b_2]$ se puede representar en la forma $S_{(b_1, b_2)} \setminus (S_{(a_1, b_2)} \cup S_{(b_1, a_2)})$.

Clasificación

1. *Vectores aleatorios discretos.* El vector aleatorio \mathbf{X} se dice *discreto* cuando existe un conjunto numerable $\mathbb{A} \subset \mathbb{R}^n$ tal que $\mathbb{P}(\mathbf{X} \in \mathbb{A}) = 1$. En tal caso, las variables aleatorias X_1, \dots, X_n son discretas y la función $p_{\mathbf{X}} : \mathbb{R}^n \rightarrow [0, 1]$ definida por

$$p_{\mathbf{X}}(\mathbf{x}) := \mathbb{P}(\mathbf{X} = \mathbf{x}) \quad (3)$$

se llama la *función de probabilidad conjunta* de \mathbf{X} . Su relación con la función de distribución conjunta es la siguiente

$$F_{\mathbf{X}}(\mathbf{x}) = \sum_{\mathbf{y} \in S_{\mathbf{x}}} p_{\mathbf{X}}(\mathbf{y}).$$

2. *Vectores aleatorios continuos.* El vector aleatorio $\mathbf{X} = (X_1, \dots, X_n)$ se dice *continuo* cuando existe una función $f_{\mathbf{X}} : \mathbb{R}^n \rightarrow \mathbb{R}^+$, llamada *densidad de probabilidades conjunta* de X_1, \dots, X_n tal que

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{S_{\mathbf{x}}} f_{\mathbf{X}}(\mathbf{y}) d\mathbf{y}.$$

(Para evitar dificultades relacionadas con el concepto de integración supondremos que las densidades son seccionalmente continuas.)

3. *Vectores aleatorios mixtos.* El vector aleatorio \mathbf{X} se dice *mixto* si no es continuo ni discreto.

Cálculo de probabilidades Dependiendo del caso, la función de probabilidad conjunta $p_{\mathbf{X}}(\mathbf{x})$, o la densidad conjunta $f_{\mathbf{X}}(\mathbf{x})$, resume toda la información relevante sobre el comportamiento del vector aleatorio \mathbf{X} . Más precisamente, para todo conjunto $A \subset \mathbb{R}^n$ “suficientemente regular”, vale que

$$\mathbb{P}(\mathbf{X} \in A) = \begin{cases} \sum_{\mathbf{x} \in A} p_{\mathbf{X}}(\mathbf{x}) & \text{en el caso discreto,} \\ \int_A f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} & \text{en el caso continuo.} \end{cases}$$

Ejemplo 1.2. Sea (X, Y) un vector aleatorio continuo con densidad conjunta $f_{X,Y}(x, y)$. Si $a < b$ y $c < d$, entonces

$$\mathbb{P}(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dx dy. \quad (4)$$

Ejemplo 1.3 (Distribución uniforme). Sea $\Lambda \subset \mathbb{R}^2$ una región acotada de área $|\Lambda|$. Si la densidad conjunta de un vector aleatorio continuo (X, Y) es de la forma

$$f_{X,Y}(x, y) = \frac{1}{|\Lambda|} \mathbf{1}\{(x, y) \in \Lambda\}, \quad (5)$$

diremos que (X, Y) está *uniformemente distribuido sobre* Λ y escribiremos $(X, Y) \sim U(\Lambda)$. Sea $\mathcal{B} \subset \Lambda$ una sub-región de Λ de área $|\mathcal{B}|$. La probabilidad de que $(X, Y) \in \mathcal{B}$ se calcula del siguiente modo

$$\mathbb{P}((X, Y) \in \mathcal{B}) = \iint_{\mathcal{B}} f_{X,Y}(x, y) dx dy = \iint_{\mathcal{B}} \frac{1}{|\Lambda|} dx dy = \frac{|\mathcal{B}|}{|\Lambda|}. \quad (6)$$

En otras palabras, la probabilidad de que $(X, Y) \in \mathcal{B}$ es la proporción del área de la región Λ contenida en la sub-región \mathcal{B} . \square

Ejemplo 1.4. Sea (X, Y) un vector aleatorio uniformemente distribuido sobre el cuadrado $[0, 1] \times [0, 1]$. ¿Cuánto vale $\mathbb{P}(XY > 1/2)$?

Debido a que el cuadrado $[0, 1] \times [0, 1]$ tiene área 1 la probabilidad requerida es el área de la región $\mathcal{B} = \{(x, y) \in [0, 1] \times [0, 1] : xy > 1/2\}$. Ahora bien,

$$(x, y) \in \mathcal{B} \iff y > 1/2x \quad (7)$$

y como $y \leq 1$, la desigualdad del lado derecho de (7) sólo es posible si $1/2 \leq x$. Vale decir,

$$\mathcal{B} = \{(x, y) : 1/2 \leq x \leq 1, 1/2x < y \leq 1\}.$$

En consecuencia,

$$\begin{aligned} \mathbb{P}(XY > 1/2) &= |\mathcal{B}| = \iint_{\mathcal{B}} 1 dx dy = \int_{1/2}^1 \left(\int_{1/(2x)}^1 1 dy \right) dx = \int_{1/2}^1 \left(1 - \frac{1}{2x} \right) dx \\ &= \frac{1}{2} + \frac{1}{2} \log\left(\frac{1}{2}\right) = \frac{1}{2}(1 - \log 2) \approx 0.1534.... \end{aligned}$$

\square

1.2. Distribuciones marginales

Sea $\mathbf{X} = (X_1, \dots, X_n)$ un vector aleatorio n -dimensional y sea $F_{\mathbf{X}}(\mathbf{x})$ su función de distribución conjunta. La coordenadas de \mathbf{X} son variables aleatorias. Cada variable individual X_i tiene su correspondiente función de distribución

$$F_{X_i}(x_i) = \mathbb{P}(X_i \leq x_i). \quad (8)$$

Para enfatizar la relación entre X_i y el vector $\mathbf{X} = (X_1, \dots, X_n)$ se dice que $F_{X_i}(x_i)$ es la *función de distribución marginal de X_i* o la *i-ésima marginal de \mathbf{X}* .

Nota Bene. Observar que, para cada $i = 1, \dots, n$, la función de distribución marginal de X_i , $F_{X_i}(x_i)$, se obtiene de la función de distribución conjunta $F_{\mathbf{X}}(x_1, \dots, x_n)$ fijando el valor de x_i y haciendo $x_j \rightarrow \infty$ para toda $j \neq i$. \square

1.2.1. Marginales discretas

Caso bidimensional. Sea (X, Y) un vector aleatorio discreto definido sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ con función de probabilidad conjunta $p_{X,Y}(x, y)$. Los números $p_{X,Y}(x, y)$, $(x, y) \in X(\Omega) \times Y(\Omega) = \{(X(\omega), Y(\omega)) : \omega \in \Omega\}$, se pueden representar en la forma de una matriz con las siguientes propiedades

$$p_{X,Y}(x, y) \geq 0, \quad \text{y} \quad \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} p_{X,Y}(x, y) = 1. \quad (9)$$

Fijando $x \in X(\Omega)$ y sumando las probabilidades que aparecen en la fila x de la matriz $p_{X,Y}(x, y)$ se obtiene

$$\sum_{y \in Y(\Omega)} p_{X,Y}(x, y) = \sum_{y \in Y(\Omega)} \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) = p_X(x). \quad (10)$$

Fijando $y \in Y(\Omega)$ y sumando las probabilidades que aparecen en la columna y de la matriz $p_{X,Y}(x, y)$ se obtiene

$$\sum_{x \in X(\Omega)} p_{X,Y}(x, y) = \sum_{x \in X(\Omega)} \mathbb{P}(X = x, Y = y) = \mathbb{P}(Y = y) = p_Y(y). \quad (11)$$

En otras palabras, sumando las probabilidades por filas obtenemos la función de probabilidad *marginal* de la variable aleatoria X y sumando las probabilidades por columnas obtenemos la función de probabilidad *marginal* de la variable aleatoria Y . El adjetivo “marginal” que reciben las funciones de probabilidad $p_X(x)$ y $p_Y(y)$ refiere a la apariencia externa que adoptan (10) y (11) en una tabla de doble entrada.

Ejemplo 1.5. En una urna hay 6 bolas rojas, 5 azules y 4 verdes. Se extraen dos. Sean X la cantidad de bolas rojas extraídas e Y la cantidad de azules.

Existen $\binom{15}{2} = 105$ resultados posibles. La cantidad de resultados con x rojas, y azules y $2 - (x + y)$ verdes es

$$\binom{6}{x} \binom{5}{y} \binom{4}{2 - (x + y)}$$

Usando esa fórmula y poniendo $q = 1/105$ obtenemos

$x \setminus y$	0	1	2	p_X
0	$6q$	$20q$	$10q$	$36q$
1	$24q$	$30q$	0	$54q$
2	$15q$	0	0	$15q$
p_Y	$45q$	$50q$	$10q$	

Figura 2: Distribución conjunta de (X, Y) . En el margen derecho de la tabla se encuentra la distribución marginal de X y en el margen inferior, la marginal de Y . \square

Caso general. Para cada $i = 1, \dots, n$, la función de probabilidad marginal de X_i , $p_{X_i}(x_i)$, se puede obtener fijando la variable x_i y sumando la función de probabilidad conjunta $p_{\mathbf{X}}(\mathbf{x})$ respecto de las demás variables

$$p_{X_i}(x_i) = \sum_{\mathbf{x}_{\{i\}^c}} p_{\mathbf{X}}(\mathbf{x}).$$

1.2.2. Marginales continuas

Sea (X, Y) un vector aleatorio continuo con función densidad conjunta $f_{X,Y}(x, y)$.

Las funciones de distribución *marginales* de las variables individuales X e Y se obtienen de la distribución conjunta haciendo lo siguiente

$$F_X(x) = \mathbb{P}(X \leq x) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y) = \int_{-\infty}^x \left(\int_{-\infty}^{\infty} f_{X,Y}(s, y) dy \right) ds, \quad (12)$$

$$F_Y(y) = \mathbb{P}(Y \leq y) = \lim_{x \rightarrow \infty} F_{X,Y}(x, y) = \int_{-\infty}^y \left(\int_{-\infty}^{\infty} f_{X,Y}(x, t) dx \right) dt. \quad (13)$$

Aplicando en (12) y en (13) el Teorema Fundamental del Cálculo Integral se obtiene que las funciones de distribución marginales $F_X(x)$ y $F_Y(y)$ son derivables (salvo quizás en un conjunto despreciable de puntos) y vale que

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad (14)$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx. \quad (15)$$

En consecuencia, las variables aleatorias X e Y son *individualmente (absolutamente) continuas* con *densidades “marginales”* $f_X(x)$ y $f_Y(y)$, respectivamente.

Ejemplo 1.6 (Distribución uniforme). Sea $\Lambda \subset \mathbb{R}^2$ una región del plano acotada, que para simplificar supondremos convexa, y sea (X, Y) un vector aleatorio uniformemente distribuido sobre Λ . La densidad marginal de X en la abscisa x es igual al cociente entre el ancho de Λ en x y el área de Λ . \square

Ejemplo 1.7 (Dardos). Consideramos un juego de dardos de blanco circular Λ de radio 1 centrado en el origen del plano: $\Lambda = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$. Un tirador lanza

un dardo al azar sobre Λ y se clava en un punto de coordenadas (X, Y) . El punto (X, Y) está uniformemente distribuido sobre Λ . Debido a que el área de Λ es igual a π , la densidad conjunta de X e Y es

$$f_{X,Y}(x, y) = \frac{1}{\pi} \mathbf{1}\{x^2 + y^2 \leq 1\}.$$

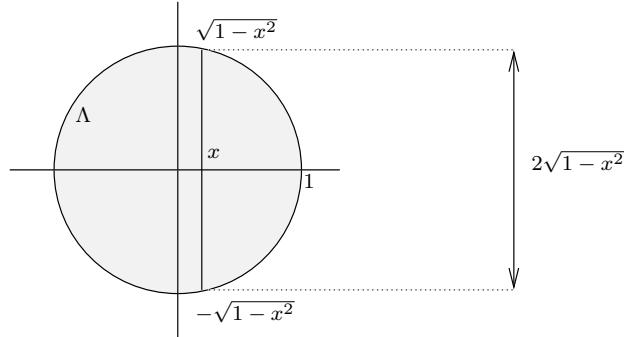


Figura 3: Para cada $x \in [-1, 1]$ se observa que el ancho del círculo en x es $2\sqrt{1 - x^2}$.

Si se observa la Figura 3 es claro que la densidad marginal de X es

$$f_X(x) = \frac{2\sqrt{1 - x^2}}{\pi} \mathbf{1}\{x \in [-1, 1]\},$$

y por razones de simetría la densidad marginal de Y debe ser

$$f_Y(y) = \frac{2\sqrt{1 - y^2}}{\pi} \mathbf{1}\{y \in [-1, 1]\}.$$

□

Caso general. Para cada $i = 1, \dots, n$, la densidad marginal de X_i , $f_{X_i}(x_i)$, se puede obtener fijando la variable x_i e integrando la densidad conjunta $f_{\mathbf{X}}(x)$ respecto de las demás variables

$$f_{X_i}(x_i) = \int_{\mathbb{R}^{n-1}} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}_{\{i\}^c}.$$

Nota Bene: Conjuntas y marginales. A veces, es necesario conocer la distribución de una sub-colección de variables aleatorias. En el caso bidimensional este problema no se manifiesta porque se reduce al cálculo de las marginales. Para cada subconjunto de índices $\Lambda \subset \{1, 2, \dots, n\}$ la función de distribución conjunta de las variables $X_i : i \in \Lambda$, $F_{\Lambda}(\mathbf{x}_{\Lambda})$, se obtiene fijando los valores de las coordenadas $x_i : i \in \Lambda$ y haciendo $x_j \rightarrow \infty$ para toda $j \notin \Lambda$.

En el caso discreto, la función de probabilidad conjunta de las variables $X_i : i \in \Lambda$, $p_{\Lambda}(x_{\Lambda})$, se obtiene fijando la variables $x_i : i \in \Lambda$ y sumando la función de probabilidad conjunta $p(\mathbf{x})$ respecto de las demás variables

$$p_{\Lambda}(\mathbf{x}_{\Lambda}) = \sum_{\mathbf{x}_{\Lambda^c}} p_{\mathbf{X}}(\mathbf{x}).$$

En el caso continuo, la densidad conjunta de las variables X_Λ , $f_\Lambda(\mathbf{x}_\Lambda)$, se obtiene fijando los valores de las variables $x_i : i \in \Lambda$ e integrando la densidad conjunta $f(x)$ respecto de las demás variables

$$f_\Lambda(\mathbf{x}_\Lambda) = \int_{\mathbb{R}^{n-m}} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}_{\Lambda^c}.$$

donde m es la cantidad de índices contenidos en el conjunto Λ .

1.3. Independencia

Las variables X_1, \dots, X_n son *independientes* si para cualquier colección de conjuntos (medibles) $A_1, \dots, A_n \subset \mathbb{R}$, los eventos $\{X_1 \in A_1\}, \dots, \{X_n \in A_n\}$ son independientes.

Tomando conjuntos de la forma $A_i = (-\infty, x_i]$ se deduce que la independencia de X_1, \dots, X_n implica

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}\left(\bigcap_{i=1}^n \{X_i \leq x_i\}\right) = \prod_{i=1}^n \mathbb{P}(X_i \leq x_i) = \prod_{i=1}^n F_{X_i}(x_i). \quad (16)$$

Dicho en palabras, la independencia de las variables implica que *su función de distribución conjunta se factoriza como el producto de todas las marginales*.

Recíprocamente, se puede demostrar que si para cada $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ se verifica la ecuación (16), las variables aleatorias X_1, \dots, X_n son independientes. (La demostración es técnica y no viene al caso). Esta equivalencia reduce al mínimo las condiciones que permiten caracterizar la independencia de variables aleatorias y motivan la siguiente definición más simple.

Definición 1.8 (Independencia de una cantidad finita de variables aleatorias). Diremos que las variables aleatorias X_1, \dots, X_n son *independientes* si la ecuación (16) se verifica en todo $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$.

Definición 1.9 (Independencia). Dada una familia de variables aleatorias $(X_i : i \in \mathbb{I})$ definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, diremos que sus variables son (*conjuntamente*) *independientes* si para cualquier subconjunto finito de índices $J \subset \mathbb{I}$ las variables $X_i, i \in J$ son independientes.

Nota Bene. La independencia de las variables aleatorias X_1, \dots, X_n es equivalente a la factorización de la distribución conjunta como producto de sus distribuciones marginales. Más aún, esta propiedad se manifiesta a nivel de la función de probabilidad, $p_{\mathbf{X}}(\mathbf{x})$ o de la densidad conjunta, $f_{\mathbf{X}}(\mathbf{x})$, del vector aleatorio $\mathbf{X} = (X_1, \dots, X_n)$, según sea el caso. Para ser más precisos, X_1, \dots, X_n son independientes si y solo si

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^n p_{X_i}(x_i) && \text{en el caso discreto,} \\ f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^n f_{X_i}(x_i) && \text{en el caso continuo.} \end{aligned}$$

Ejemplo 1.10 (Números al azar). Se elige al azar un número U del intervalo $[0, 1]$. Sea $U = 0.X_1X_2X_3\dots$ el desarrollo decimal de U . Mostraremos que los dígitos de U son independientes entre sí y que cada uno de ellos se distribuye uniformemente sobre el conjunto $\{0, 1, \dots, 9\}$.

El problema se reduce a mostrar que para cada $n \geq 2$ las variables aleatorias X_1, X_2, \dots, X_n son independientes entre sí y que para cada $k \geq 1$ y todo $x_k \in \{0, 1, \dots, 9\}$, $\mathbb{P}(X_k = x_k) = 1/10$.

Primero observamos que para cada $n \geq 1$ y para todo $(x_1, \dots, x_n) \in \{0, 1, \dots, 9\}^n$ vale que

$$\bigcap_{i=1}^n \{X_i = x_i\} \iff U \in \left[\sum_{i=1}^n \frac{x_i}{10^i}, \sum_{i=1}^n \frac{x_i}{10^i} + \frac{1}{10^n} \right).$$

En consecuencia,

$$\mathbb{P}\left(\bigcap_{i=1}^n \{X_i = x_i\}\right) = \frac{1}{10^n}. \quad (17)$$

Para calcular las marginales de los dígitos observamos que para cada $x_k \in \{0, 1, \dots, 9\}$ vale que

$$\{X_k = x_k\} = \bigcup_{(x_1, \dots, x_{k-1}) \in \{0, 1, \dots, 9\}^{k-1}} \left[\left(\bigcap_{i=1}^{k-1} \{X_i = x_i\} \right) \cap \{X_k = x_k\} \right].$$

De acuerdo con (17) cada uno de los 10^{k-1} eventos que aparecen en la unión del lado derecho de la igualdad tiene probabilidad $1/10^k$ y como son disjuntos dos a dos obtenemos que

$$\mathbb{P}(X_k = x_k) = 10^{k-1} \frac{1}{10^k} = \frac{1}{10}. \quad (18)$$

De (17) y (18) se deduce que para todo $(x_1, \dots, x_n) \in \{0, 1, \dots, 9\}^n$ vale que

$$\mathbb{P}\left(\bigcap_{i=1}^n \{X_i = x_i\}\right) = \prod_{i=1}^n \mathbb{P}(X_i = x_i).$$

Por lo tanto, las variables aleatorias X_1, X_2, \dots, X_n son independientes entre sí y cada una de ellas se distribuye uniformemente sobre el conjunto $\{0, 1, \dots, 9\}$. \square

1.3.1. Caso bidimensional discreto

Sea (X, Y) un vector aleatorio discreto con función de probabilidad conjunta $p_{X,Y}(x, y)$ y marginales $p_X(x)$ y $p_Y(y)$. Las variables X, Y son *independientes* si para cada pareja de valores $x \in X(\Omega)$, $y \in Y(\Omega)$ vale que

$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \quad (19)$$

En otras palabras, la matriz $p_{X,Y}(x, y)$ es la tabla de multiplicar de las marginales $p_X(x)$ y $p_Y(y)$.

Ejemplo 1.11. Se arrojan dos dados equilibrados y se observan las variables aleatorias X e Y definidas por $X = \text{"el resultado del primer dado"}$ e $Y = \text{"el mayor de los dos resultados"}$.

El espacio de muestral asociado al experimento se puede representar en la forma $\Omega = \{1, 2, \dots, 6\}^2$, cada punto $(i, j) \in \Omega$ indica que el resultado del primer dado es i y el resultado del segundo es j . Para reflejar que arrojamos *dos dados equilibrados*, todos los puntos de Ω serán equiprobables, i.e., para cada $(i, j) \in \Omega$ se tiene $\mathbb{P}(i, j) = 1/36$. Formalmente las variables aleatorias X e Y están definidas por

$$X(i, j) := i, \quad Y(i, j) := \max\{i, j\}. \quad (20)$$

Distribución conjunta y distribuciones marginales de X e Y . En primer lugar vamos a representar el espacio muestral Ω en la forma de una matriz para poder observar más claramente los resultados posibles

$$\begin{pmatrix} (1, 1) & (1, 2) & (1, 3) & (1, 4) & (1, 5) & (1, 6) \\ (2, 1) & (2, 2) & (2, 3) & (2, 4) & (2, 5) & (2, 6) \\ (3, 1) & (3, 2) & (3, 3) & (3, 4) & (3, 5) & (3, 6) \\ (4, 1) & (4, 2) & (4, 3) & (4, 4) & (4, 5) & (4, 6) \\ (5, 1) & (5, 2) & (5, 3) & (5, 4) & (5, 5) & (5, 6) \\ (6, 1) & (6, 2) & (6, 3) & (6, 4) & (6, 5) & (6, 6) \end{pmatrix}$$

Figura 4: Resultados posibles del experimento aleatorio que consiste en arrojar dos dados.

Debido a que $Y \geq X$, tenemos que $p_{X,Y}(x, y) = 0$ para todo $1 \leq y < x \leq 6$. En los otros casos, i.e., $1 \leq x \leq y \leq 6$, para calcular el valor de $p_{X,Y}(x, y)$ hay que contar la cantidad de elementos de la fila x , de la matriz representada en la Figura 4, que contengan alguna coordenada igual a y . Multiplicando por $q = \frac{1}{36}$ la cantidad encontrada se obtiene $p_{X,Y}(x, y)$. En la figura 5 representamos la distribución conjunta $p_{X,Y}(x, y)$ y las distribuciones marginales p_X y p_Y .

$x \setminus y$	1	2	3	4	5	6	p_X
1	q	q	q	q	q	q	$6q$
2	0	$2q$	q	q	q	q	$6q$
3	0	0	$3q$	q	q	q	$6q$
4	0	0	0	$4q$	q	q	$6q$
5	0	0	0	0	$5q$	q	$6q$
6	0	0	0	0	0	$6q$	$6q$
p_Y	q	$3q$	$5q$	$7q$	$9q$	$11q$	

Figura 5: Distribución conjunta de (X, Y) . En el margen derecho se encuentra la distribución marginal de X y en el margen inferior, la marginal de Y . Para abreviar hemos puesto $q = \frac{1}{36}$.

De acuerdo con los resultados expuestos en la tabla que aparece en la Figura 5, las distribuciones marginales son

$$p_X(x) = \frac{1}{6}, \quad p_Y(y) = \frac{2y - 1}{36}.$$

Debido a que no se trata de una tabla de multiplicar las variables X e Y no son independientes. Lo que, por otra parte, constituye una obviedad.

Criterio para detectar dependencia. Cuando en la tabla de la distribución conjunta de dos variables hay un 0 ubicado en la intersección de una fila y una columna de sumas positivas, las variables no pueden ser independientes. (Las variables del Ejemplo 1.5 no son independientes.) \square

1.3.2. Caso bidimensional continuo

Sean X e Y variables aleatorias con densidad conjunta $f_{X,Y}(x,y)$ y marginales $f_X(x)$ y $f_Y(y)$. Las variables aleatorias X e Y son *independientes* si y solo si

$$f_{X,Y}(x,y) = f_X(x)f_Y(y). \quad (21)$$

En otras palabras, X e Y son independientes si y solo si su densidad conjunta se factoriza como el producto de las marginales.

Criterios para detectar (in)dependencia.

1. La independencia de X e Y equivale a la existencia de dos funciones $f_1(x)$ y $f_2(y)$ tales que $f_{X,Y}(x,y) = f_1(x)f_2(y)$. Por lo tanto, para verificar independencia basta comprobar que la densidad conjunta se puede factorizar como *alguna* función de x por *alguna* función de y , siendo innecesario verificar que se trata de las densidades marginales. (*Ejercicio*)

2. La factorización (21) implica que, si X e Y son independientes, el recinto del plano $Sop(f_{X,Y}) := \{(x,y) \in \mathbb{R}^2 : f_{X,Y}(x,y) > 0\}$, llamado *el soporte de la densidad conjunta* $f_{X,Y}$, debe coincidir con el producto cartesiano de los soportes de sus densidades marginales: $Sop(f_X) \times Sop(f_Y) = \{x \in \mathbb{R} : f_X(x) > 0\} \times \{y \in \mathbb{R} : f_Y(y) > 0\}$. Por ejemplo, si el soporte de la densidad conjunta es *conexo* y no es un rectángulo las variables X e Y no pueden ser independientes. (Ver el Ejemplo 1.7.)

Ejemplo 1.12. Sean X e Y variables aleatorias independientes con distribución uniforme sobre el intervalo $(0, L)$. Una vara de longitud L metros se quiebra en dos puntos cuyas distancias a una de sus puntas son X e Y metros. Calcular la probabilidad de que las tres piezas se puedan usar para construir un triángulo.

Primero designamos mediante L_1 , L_2 y L_3 a las longitudes de las tres piezas. Las tres piezas se pueden usar para construir un triángulo si y solamente si se satisfacen las desigualdades triangulares

$$L_1 + L_2 > L_3, \quad L_1 + L_3 > L_2 \quad \text{y} \quad L_2 + L_3 > L_1. \quad (22)$$

Vamos a distinguir dos casos: el caso en que $X \leq Y$ y el caso en que $Y < X$. En el primer caso, $X \leq Y$, tenemos que $L_1 = X$, $L_2 = Y - X$ y $L_3 = L - Y$ y las desigualdades triangulares (22) son equivalentes a las siguientes

$$Y > L/2, \quad X + L/2 > Y \quad \text{y} \quad L/2 > X. \quad (23)$$

En el segundo caso, $Y < X$, tenemos que $L_1 = Y$, $L_2 = X - Y$ y $L_3 = L - X$ y las desigualdades triangulares (22) son equivalentes a las siguientes

$$X > L/2, \quad Y > X - L/2 \quad \text{y} \quad L/2 > Y. \quad (24)$$

Por lo tanto, las tres piezas se pueden usar para construir un triángulo si y solamente si $(X, Y) \in \mathcal{B}$, donde

$$\begin{aligned}\mathcal{B} = & \{(x, y) \in (0, L) \times (0, L) : 0 < x < L/2, L/2 < y < x + L/2\} \\ & \cup \{(x, y) \in (0, L) \times (0, L) : L/2 < x < L, x - L/2 < y < L/2\}. \end{aligned} \quad (25)$$

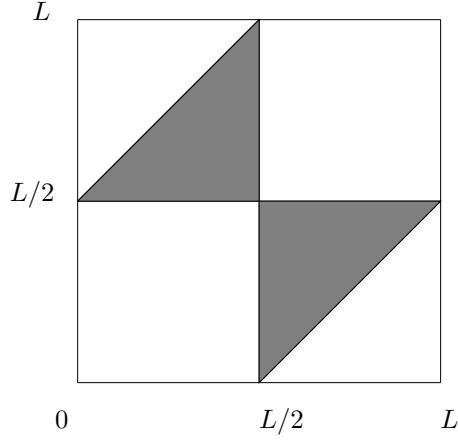


Figura 6: La región sombreada representa al conjunto \mathcal{B} que es la unión de dos triángulos disjuntos cada uno de área $L^2/8$.

La hipótesis de que X e Y son independientes con distribución uniforme sobre el intervalo $(0, L)$ significa que $(X, Y) \sim \mathcal{U}(\Lambda)$, donde Λ es el cuadrado de lado $(0, L)$

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) = \left(\frac{1}{L}\mathbf{1}\{0 < x < L\}\right)\left(\frac{1}{L}\mathbf{1}\{0 < y < L\}\right) = \frac{1}{L^2}\mathbf{1}\{(x, y) \in \Lambda\}.$$

De (6) se deduce que

$$\mathbb{P}((X, Y) \in \mathcal{B}) = \frac{|\mathcal{B}|}{|\Lambda|} = \frac{(2/8)L^2}{L^2} = \frac{1}{4}. \quad (26)$$

□

2. Bibliografía consultada

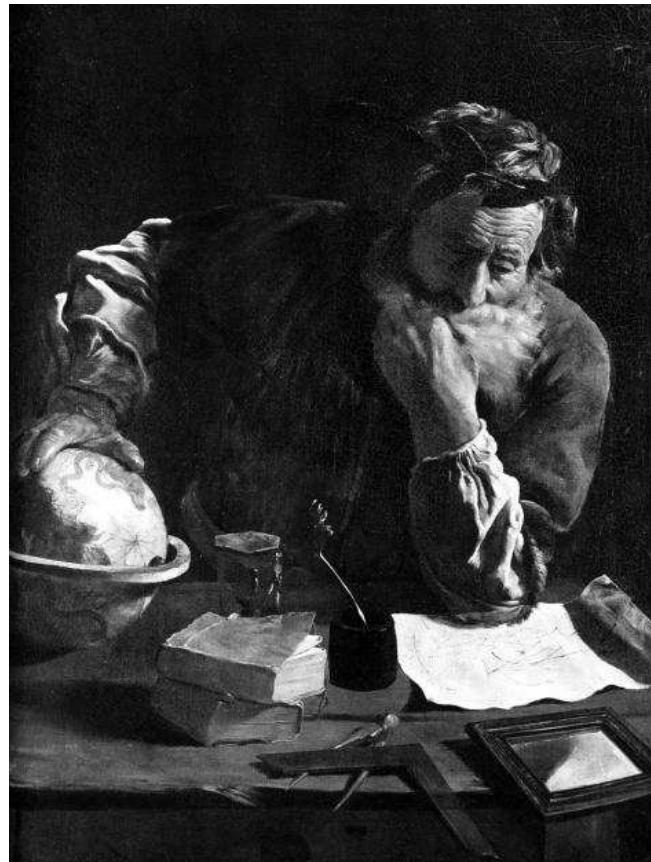
Para redactar estas notas se consultaron los siguientes libros:

1. Bertsekas, D. P., Tsitsiklis, J. N.: Introduction to Probability. M.I.T. Lecture Notes. (2000)
2. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons, New York. (1968)
3. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 2. John Wiley & Sons, New York. (1971)
4. Ross, S.: Introduction to Probability Models. Academic Press, San Diego. (2007)

Variables aleatorias: momentos (Borradores, Curso 23)

Sebastian Grynberg

27 de marzo 2013



Denme un punto de apoyo y moveré el mundo
(Arquímedes de Siracusa)

Índice

1. Esperanza	2
1.1. Definición	3
1.2. Cálculo	8
1.3. Propiedades	10
1.4. Dividir y conquistar	11
2. Varianza	12
2.1. Definición	12
2.2. Cálculo	13
2.3. Propiedades	14
3. Covarianza	14
3.1. Definición	14
3.2. Cálculo	14
3.3. Propiedades	16
3.4. Varianza de sumas	16
4. Algunas desigualdades	17
4.1. Cauchy-Schwartz	17
4.2. Chebyshev	18
5. La ley débil de los grandes números	20
6. Distribuciones particulares	22
7. Bibliografía consultada	28

1. Esperanza

La información relevante sobre el comportamiento de una variable aleatoria está contenida en su función de distribución. Sin embargo, en la práctica, es útil disponer de algunos números representativos de la variable aleatoria que resuman esa información.

Motivación Se gira una rueda de la fortuna varias veces. En cada giro se puede obtener alguno de los siguientes números x_1, x_2, \dots, x_k -que representan la cantidad de dinero que se obtiene en el giro- con probabilidades $p(x_1), p(x_2), \dots, p(x_k)$, respectivamente. ¿Cuánto dinero se “espera” obtener como recompensa “por cada giro”? Los términos “espera” y “por cada giro” son un tanto ambiguos, pero se pueden interpretar de la siguiente manera.

Si la rueda se gira n veces y $n(x_i)$ es la cantidad de veces que se obtiene x_i , la cantidad total de dinero recibida es $\sum_{i=1}^k n(x_i)x_i$ y la cantidad media por giro es $\mu = \frac{1}{n} \sum_{i=1}^k n(x_i)x_i$. Interpretando las probabilidades como frecuencias relativas obtenemos que para n suficientemente grande la cantidad de dinero que se “espera” recibir “por cada giro” es

$$\mu = \frac{1}{n} \sum_{i=1}^k x_i n(x_i) = \sum_{i=1}^k x_i \frac{n(x_i)}{n} \approx \sum_{i=1}^k x_i p(x_i).$$

1.1. Definición

Definición 1.1 (Esperanza de una variable discreta). Sea X una variable aleatoria discreta. La *esperanza* de X , denotada por $\mathbb{E}[X]$, es el promedio ponderado

$$\mathbb{E}[X] := \sum_{x \in \mathbb{A}} x \mathbb{P}(X = x), \quad (1)$$

donde $\mathbb{A} = \{x \in \mathbb{R} : F(x) - F(x-) > 0\}$ es el conjunto de todos los átomos de la función distribución de X .

Ejemplo 1.2 (Esperanza de la función indicadora). Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad. Para cualquier evento $A \in \mathcal{A}$ vale que

$$\mathbb{E}[\mathbf{1}\{\omega \in A\}] = 0 \cdot (1 - \mathbb{P}(A)) + 1 \cdot \mathbb{P}(A) = \mathbb{P}(A). \quad (2)$$

□

La esperanza como centro de gravedad. La noción de esperanza es análoga a la noción de centro de gravedad para un sistema de partículas discreto.

Se consideran n partículas ubicadas en los puntos x_1, \dots, x_n cuyos pesos respectivos son $p(x_1), \dots, p(x_n)$. No se pierde generalidad si se supone que $\sum_{i=1}^n p(x_i) = 1$. El centro de gravedad, c , del sistema es el punto respecto de la cual la suma de los momentos causados por los pesos $p(x_i)$ es nula. Observando que

$$\sum_{i=1}^k (x_i - c) p(x_i) = 0 \iff c = \sum_{i=1}^k x_i p(x_i)$$

resulta que el centro de gravedad del sistema coincide con la esperanza de una variable aleatoria X a valores en $\{x_1, \dots, x_n\}$ tal que $\mathbb{P}(X = x_i) = p(x_i)$. □

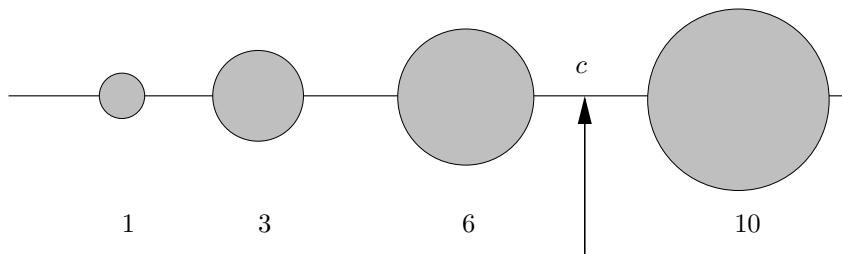


Figura 1: Interpretación de la esperanza como centro de gravedad. Se considera un sistema de cuatro “partículas” de pesos p_i proporcionales a las áreas de los círculos de radio $1/3, 2/3, 3/3, 4/3$ centrados en los puntos $x_i = 1, 3, 6, 10$, respectivamente. No se pierde generalidad si se supone que el peso total del sistema es la unidad. El centro de gravedad del sistema se encuentra en el punto $c = \sum_{i=1}^4 x_i p_i = 227/30 = 7.56\dots$

La esperanza como promedio. Sea X una variable aleatoria a valores x_1, \dots, x_n con función de probabilidades

$$\mathbb{P}(X = x) = \frac{1}{n} \mathbf{1}\{x \in \{x_1, \dots, x_n\}\}.$$

Conforme a la Definición 1.1 la esperanza de X es

$$\mathbb{E}[X] = \sum_{i=1}^n x_i \mathbb{P}(X = x_i) = \frac{1}{n} \sum_{i=1}^n x_i. \quad (3)$$

Dicho en palabras: la esperanza de una variable aleatoria uniformemente distribuida sobre los valores x_1, x_2, \dots, x_n coincide con el promedio de dichos valores. \square

Ejemplo 1.3 (Dado equilibrado). Sea X el resultado del lanzamiento de un dado equilibrado. De acuerdo con (3) la esperanza de X es

$$\mathbb{E}[X] = \frac{1}{6} \sum_{x=1}^6 x = \frac{21}{6} = \frac{7}{2}.$$

\square

Ejemplo 1.4 (Uniforme sobre el “intervalo” $\{1, 2, \dots, n\}$). La variable aleatoria del Ejemplo 1.3 es un caso particular de una variable aleatoria discreta X uniformemente distribuida sobre el “intervalo” de números enteros $\{1, 2, \dots, n\}$. De acuerdo con (3) la esperanza de X es

$$\mathbb{E}[X] = \frac{1}{n} \sum_{x=1}^n x = \frac{1}{n} \left(\frac{n(n+1)}{2} \right) = \frac{1+n}{2}.$$

\square

Ejemplo 1.5 (Moneda equilibrada). Sea N la cantidad de veces que debe lanzarse una moneda equilibrada hasta que salga cara. N es una variable aleatoria discreta a valores $1, 2, \dots$ tal que $\mathbb{P}(N = n) = (1/2)^n$, $n = 1, 2, \dots$. De acuerdo con la definición 1.1, la esperanza de N es

$$\mathbb{E}[N] = \sum_{n=1}^{\infty} n \mathbb{P}(N = n) = \sum_{n=1}^{\infty} n \left(\frac{1}{2} \right)^n.$$

Derivando ambos lados de la igualdad $\sum_{n=0}^{\infty} x^n = (1-x)^{-1}$, que vale para $|x| < 1$, se deduce que $\sum_{n=0}^{\infty} nx^{n-1} = (1-x)^{-2}$ y de allí resulta que $\sum_{n=1}^{\infty} nx^n = x(1-x)^{-2}$. Evaluando en $x = 1/2$ se obtiene que

$$\mathbb{E}[N] = \sum_{n=1}^{\infty} n \left(\frac{1}{2} \right)^n = \left(\frac{1}{2} \right) \left(\frac{1}{2} \right)^{-2} = 2.$$

\square

La noción de esperanza se extiende a *variables aleatorias absolutamente continuas* cambiando en (1) la suma por la integral y la función de probabilidades $P(X = x)$, $x \in \mathbb{A}$, por la densidad de probabilidades de la variable X .

Definición 1.6 (Esperanza de una variable absolutamente continua). Sea X una variable aleatoria absolutamente continua con densidad de probabilidades $f_X(x)$. La *esperanza de X* , denotada por $\mathbb{E}[X]$, se define por

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} xf_X(x)dx. \quad (4)$$

Ejemplo 1.7 (Fiabilidad). Sea T el tiempo de espera hasta que ocurre la primer falla en un sistema electrónico con función intensidad de fallas de la forma $\lambda(t) = 2t\mathbf{1}\{t > 0\}$. La función de distribución de T es $F_T(t) = (1 - \exp(-t^2))\mathbf{1}\{t > 0\}$. En consecuencia, T es una variable aleatoria absolutamente continua con densidad de probabilidad $f_T(t) = 2t\exp(-t^2)\mathbf{1}\{t > 0\}$. De acuerdo con la definición 1.6, la esperanza de T es

$$\mathbb{E}[T] = \int_{-\infty}^{\infty} tf_T(t)dt = \int_0^{\infty} t2t\exp(-t^2)dt = \int_0^{\infty} \exp(-t^2)dt = \frac{\sqrt{\pi}}{2}.$$

La tercera igualdad se deduce de la fórmula de integración por partes aplicada a $u = t$ y $v' = 2t\exp(-t^2)$ y la cuarta se deduce de la identidad $\int_0^{\infty} \exp(-x^2/2)dx = \sqrt{2\pi}/2$ mediante el cambio de variables $t = x/\sqrt{2}$. \square

Extendiendo la noción a variables mixtas. La noción de esperanza para variables mixtas se obtiene combinando las nociones anteriores.

Definición 1.8 (Esperanza de una variable mixta). Sea X una variable aleatoria mixta con función de distribución $F_X(x)$. La *esperanza de X* , denotada por $\mathbb{E}[X]$, se define de la siguiente manera:

$$\mathbb{E}[X] := \sum_{x \in \mathbb{A}} x\mathbb{P}(X = x) + \int_{-\infty}^{\infty} xF'_X(x)dx, \quad (5)$$

donde $\mathbb{A} = \{x \in \mathbb{R} : F_X(x) - F_X(x-) > 0\}$ es el conjunto de todos los átomos de $F_X(x)$ y $F'_X(x)$ es una función que coincide con la derivada de $F_X(x)$ en todos los puntos donde esa función es derivable y vale 0 en otro lado.

Ejemplo 1.9 (Mixtura). Sea X una variable aleatoria mixta cuya función de distribución es $F_X(x) = (\frac{2x+5}{8})\mathbf{1}\{-1 \leq x < 1\} + \mathbf{1}\{x \geq 1\}$. De acuerdo con la fórmula (5), la esperanza de X es

$$\mathbb{E}[X] = -1 \cdot \mathbb{P}(X = -1) + 1 \cdot \mathbb{P}(X = 1) + \int_{-1}^1 F'_X(x)dx = -\frac{3}{8} + \frac{1}{8} + \int_{-1}^1 \frac{2}{8}dx = \frac{1}{4}.$$

\square

Nota Bene. En todas las definiciones anteriores, se presupone que las series y/o integrales involucradas son absolutamente convergentes.

Ejemplo 1.10 (Distribución de Cauchy). Sea X una variable aleatoria con *distribución de Cauchy*. Esto es, X es absolutamente continua y admite una densidad de probabilidades de la forma

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Debido a que

$$\int_{-\infty}^{\infty} |x| f(x) dx = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)} dx = \infty,$$

X no tiene esperanza. \square

Teorema 1.11. Sea X una variable aleatoria no negativa (i.e., $F_X(x) = \mathbb{P}(X \leq x) = 0$ para todo $x < 0$). Vale que

$$\mathbb{E}[X] = \int_0^{\infty} [1 - F_X(x)] dx. \quad (6)$$

Demostración. El argumento principal está contenido en la Figura 2. El caso general se deduce usando técnicas de “paso al límite”.

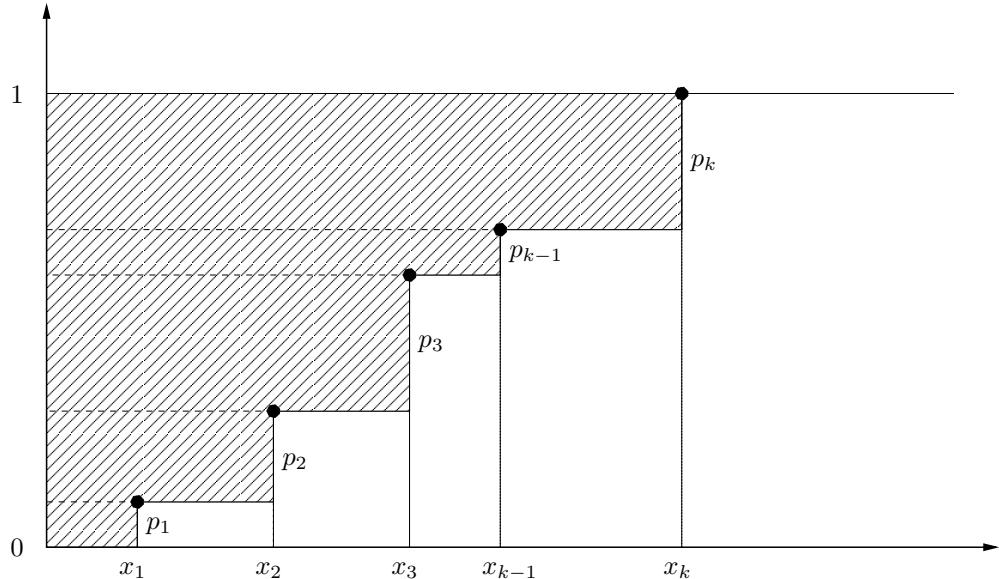


Figura 2: Argumento geométrico que muestra la validez de la identidad (6) en el caso en que X es no negativa, discreta y a valores $0 \leq x_1 < x_2 < \dots < x_k$. Si $p_i = \mathbb{P}(X = x_i)$, el área de la región sombreada es la suma $x_1p_1 + \dots + x_kp_k = \mathbb{E}[X]$ de las áreas de los rectángulos horizontales y coincide con la integral de la altura $\mathbb{P}(X > x)$. \square

Corolario 1.12. Sea X una variable aleatoria con función de distribución $F_X(x)$. Vale que

$$\mathbb{E}[X] = \int_0^{\infty} [1 - F_X(x)] dx - \int_{-\infty}^0 F_X(x) dx. \quad (7)$$

Demostración. Ejercicio. \square

Nota Bene. Las identidades (6) y (7) son interesantes porque muestran que para calcular la esperanza de una variable aleatoria basta conocer su función de distribución. De hecho, la identidad (7) ofrece una definición alternativa y unificada de la noción de esperanza. \square

Ejemplo 1.13. Una máquina fue diseñada para prestar servicios en una instalación productiva. La máquina se enciende al iniciar la jornada laboral y se apaga al finalizar la misma. Si durante ese período la máquina falla, se la repara y en esa tarea se consume el resto de la jornada.

Suponiendo que la función intensidad de fallas de la máquina es una constante $\lambda > 0$ (y que el tiempo se mide en jornadas laborales), hallar el máximo valor de λ que permita asegurar con una probabilidad mayor o igual que $2/3$ que la máquina prestará servicios durante una jornada laboral completa. Para ese valor de λ , hallar (y graficar) la función de distribución del tiempo, T , de funcionamiento de la máquina durante una jornada laboral y calcular el tiempo medio de funcionamiento, $\mathbb{E}[T]$.

Solución. Si T_1 es el tiempo que transcurre desde que se enciende la máquina hasta que ocurre la primer falla, el evento “la máquina funciona durante una jornada laboral completa” se describe mediante $\{T_1 > 1\}$. Queremos hallar el máximo $\lambda > 0$ tal que $\mathbb{P}(T_1 > 1) \geq 2/3$. Debido a que la función intensidad de fallas es una constante λ se tiene que $\mathbb{P}(T_1 > t) = e^{-\lambda t}$. En consecuencia, $\mathbb{P}(T_1 > 1) \geq 2/3 \iff e^{-\lambda} \geq 2/3 \iff \lambda \leq -\log(2/3)$. Por lo tanto, $\lambda = -\log(2/3)$. En tal caso, $\mathbb{P}(T_1 > 1) = 2/3$.

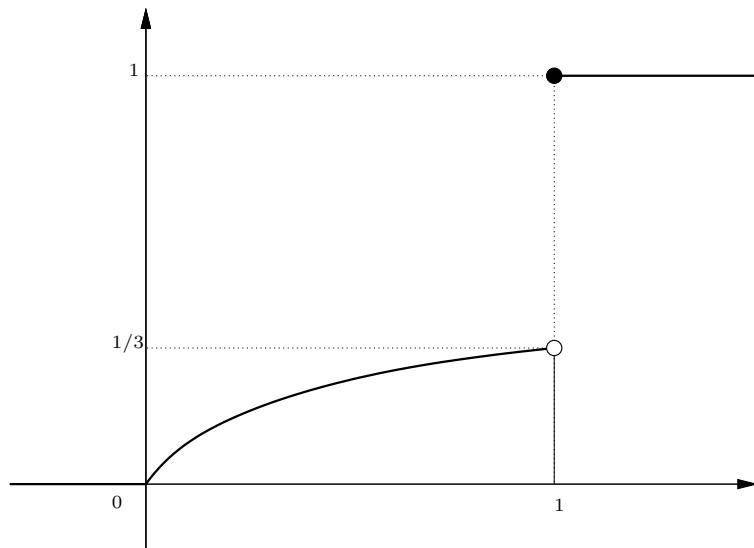


Figura 3: Gráfico de la función de distribución de T .

El tiempo de funcionamiento de la máquina por jornada laboral es $T = \min\{T_1, 1\}$. Para $t > 0$ vale que

$$\begin{aligned} F_T(t) &= \mathbb{P}(T \leq t) = 1 - \mathbb{P}(T > t) = 1 - \mathbb{P}(\min\{T_1, 1\} > t) \\ &= 1 - \mathbb{P}(T_1 > t)\mathbf{1}\{1 > t\} = 1 - e^{\log(2/3)t}\mathbf{1}\{t < 1\} \\ &= \left(1 - e^{\log(2/3)t}\right)\mathbf{1}\{0 \leq t < 1\} + \mathbf{1}\{t \geq 1\}. \end{aligned}$$

Como $T > 0$ y conocemos la función $\mathbb{P}(T > t)$ lo más sencillo para calcular la esperanza es usar la fórmula $\mathbb{E}[T] = \int_0^\infty \mathbb{P}(T > t)dt$:

$$\begin{aligned}\mathbb{E}[T] &= \int_0^\infty \mathbb{P}(T > t)dt = \int_0^1 e^{\log(2/3)t} dt = \frac{e^{\log(2/3)t}}{\log(2/3)} \Big|_0^1 = \frac{2/3 - 1}{\log(2/3)} \\ &= \frac{-1/3}{\log(2/3)} \approx 0.822...\end{aligned}$$

□

1.2. Cálculo

Sea X una variable aleatoria cuya función de distribución conocemos. Queremos calcular la esperanza de alguna función de X , digamos, $g(X)$. ¿Cómo se puede efectuar ese cálculo? Una manera es la siguiente: (1) Hallamos la función de distribución de la variable aleatoria $Y = g(X)$ a partir del conocimiento que tenemos sobre la distribución de X :

$$F_Y(y) := \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \in g^{-1}(-\infty, y]).$$

(2) Usando la distribución de Y calculamos la esperanza $\mathbb{E}[g(X)] = \mathbb{E}[Y]$ por definición.

Ejemplo 1.14. Sea X una variable aleatoria discreta tal que $\mathbb{P}(X = 0) = 0.2$, $\mathbb{P}(X = 1) = 0.5$ y $\mathbb{P}(X = 2) = 0.3$. Queremos calcular $\mathbb{E}[X^2]$. Poniendo $Y = X^2$ obtenemos una variable aleatoria a valores en $\{0^2, 1^2, 2^2\}$ tal que $\mathbb{P}(Y = 0) = 0.2$ $\mathbb{P}(Y = 1) = 0.5$ y $\mathbb{P}(Y = 4) = 0.3$. Por definición, $\mathbb{E}[X^2] = \mathbb{E}[Y] = 0(0.2) + 1(0.5) + 4(0.3) = 1.7$. □

Ejemplo 1.15. Sea X una variable aleatoria con distribución uniforme sobre el intervalo $(0, 1)$. Queremos calcular $\mathbb{E}[X^3]$. Ponemos $Y = X^3$ y calculamos su función de distribución: para cada $0 < y < 1$ vale que $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^3 \leq y) = \mathbb{P}(X \leq y^{1/3}) = y^{1/3}$. Derivando $F_Y(y)$ obtenemos la densidad de probabilidad de Y : $f_Y(y) = \frac{1}{3}y^{-2/3}\mathbf{1}\{0 < y < 1\}$. Por definición,

$$\mathbb{E}[X^3] = \mathbb{E}[Y] = \int_{-\infty}^\infty y f_Y(y) dy = \int_0^1 y \frac{1}{3}y^{-2/3} dy = \frac{1}{3} \int_0^1 y^{1/3} dy = \frac{1}{3} \frac{3}{4} y^{4/3} \Big|_0^1 = \frac{1}{4}.$$

□

Nota Bene. Existe una manera mucho más simple para calcular la esperanza de $Y = g(X)$ que no recurre al procedimiento de determinar primero la distribución de Y para luego calcular su esperanza por definición. El Teorema siguiente muestra cómo hacerlo.

Teorema 1.16. Sea X una variable aleatoria y sea $g : \mathbb{R} \rightarrow \mathbb{R}$ una función tal que $g(X)$ también es una variable aleatoria.

(a) Si X es discreta con átomos en el conjunto \mathbb{A} , entonces

$$\mathbb{E}[g(X)] = \sum_{x \in \mathbb{A}} g(x) \mathbb{P}(X = x). \quad (8)$$

(b) Si X es continua con densidad de probabilidad $f_X(x)$ y $g(X)$ es continua, entonces

$$\mathbb{E}[g(X)] = \int_{-\infty}^\infty g(x) f_X(x) dx. \quad (9)$$

(c) Si X es mixta,

$$\mathbb{E}[g(X)] = \sum_{x \in \mathbb{A}} g(x)\mathbb{P}(X = x) + \int_{-\infty}^{\infty} g(x)F'_X(x)dx, \quad (10)$$

donde \mathbb{A} es el conjunto de todos los átomos de $F_X(x)$ y $F'_X(x)$ es un función que coincide con la derivada de $F_X(x)$ en todos los puntos donde esa función es derivable y vale cero en otro lado.

Demostración. Para simplificar la demostración supondremos que $g \geq 0$.

(a) Por el Teorema 1.11 tenemos que

$$\begin{aligned} \mathbb{E}[g(X)] &= \int_0^\infty \mathbb{P}(g(X) > y)dy = \int_0^\infty \left(\sum_{x \in \mathbb{A}} \mathbf{1}\{g(x) > y\}\mathbb{P}(X = x) \right) dy \\ &= \sum_{x \in \mathbb{A}} \left(\int_0^\infty \mathbf{1}\{g(x) > y\}dy \right) \mathbb{P}(X = x) = \sum_{x \in \mathbb{A}} g(x)\mathbb{P}(X = x). \end{aligned}$$

(b) Por el Teorema 1.11 tenemos que

$$\begin{aligned} \mathbb{E}[g(X)] &= \int_0^\infty \mathbb{P}(g(X) > y)dy = \int_0^\infty \left(\int_{\{x: g(x) > y\}} f(x)dx \right) dy \\ &= \int_{-\infty}^\infty \left(\int_0^{g(x)} dy \right) f(x)dx = \int_{-\infty}^\infty g(x)f(x)dx. \end{aligned}$$

(c) Se obtiene combinando adecuadamente los resultados (a) y (b). □

Ejemplo 1.17. Aplicando la parte (a) del Teorema 1.16 al Ejemplo 1.14 se obtiene

$$\mathbb{E}[X^2] = 0^2(0.2) + 1^2(0.5) + 2^2(0.3) = 1.7.$$

□

Ejemplo 1.18. Aplicando la parte (b) del Teorema 1.16 al Ejemplo 1.15 se obtiene

$$\mathbb{E}[X^3] = \int_0^1 x^3 dx = \frac{1}{4}.$$

□

Teorema 1.19 (Cálculo de Esperanzas). Sea \mathbf{X} un vector aleatorio y sea $g : \mathbb{R}^n \rightarrow \mathbb{R}$ una función tal que $g(\mathbf{X})$ es una variable aleatoria. Si la variable aleatoria $g(\mathbf{X})$ tiene esperanza finita, entonces

$$\mathbb{E}[g(\mathbf{X})] = \begin{cases} \sum_{\mathbf{x}} g(\mathbf{x})p_{\mathbf{X}}(\mathbf{x}) & \text{en el caso discreto,} \\ \int_{\mathbb{R}^n} g(\mathbf{x})f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} & \text{en el caso continuo,} \end{cases}$$

donde, según sea el caso, $p_{\mathbf{X}}(\mathbf{x})$ y $f_{\mathbf{X}}(\mathbf{x})$ son la función de probabilidad y la densidad conjunta del vector \mathbf{X} , respectivamente.

Demostración. Enteramente análoga a la que hicimos en dimensión 1. \square

Sobre el cálculo de esperanzas. El Teorema 1.19 es una herramienta práctica para calcular esperanzas. Su resultado establece que si queremos calcular la esperanza de una transformación unidimensional del vector \mathbf{X} , $g(\mathbf{X})$, no necesitamos calcular la distribución de $g(\mathbf{X})$. La esperanza $\mathbb{E}[g(\mathbf{X})]$ puede calcularse directamente a partir del conocimiento de la distribución conjunta de \mathbf{X} . \square

Corolario 1.20 (Esperanza de las marginales). Sea $\mathbf{X} = (X_1, \dots, X_n)$ un vector aleatorio. Si la variable X_i tiene esperanza finita, entonces

$$\mathbb{E}[X_i] = \begin{cases} \sum_{\mathbf{x}} x_i p_{\mathbf{X}}(\mathbf{x}) & \text{en el caso discreto,} \\ \int_{\mathbb{R}^n} x_i f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} & \text{en el caso continuo.} \end{cases}$$

1.3. Propiedades

- (a) Si $X = 1$, entonces $\mathbb{E}[X] = 1$.
- (b) *Monotonía.* Si X_1 y X_2 son dos variables aleatorias tales que $X_1 \leq X_2$, entonces $\mathbb{E}[X_1] \leq \mathbb{E}[X_2]$.
- (c) Si X es una variable aleatoria tal que $\mathbb{E}[X^n]$ es finita y a_0, a_1, \dots, a_n son constantes, entonces

$$\mathbb{E} \left[\sum_{k=0}^n a_k X^k \right] = \sum_{k=0}^n a_k \mathbb{E}[X^k]. \quad (11)$$

- (d) *Linealidad.* Si las variables aleatorias X_1, \dots, X_n tienen esperanza finita y a_1, a_2, \dots, a_n son constantes, entonces

$$\mathbb{E} \left[\sum_{i=1}^n a_i X_i \right] = \sum_{i=1}^n a_i \mathbb{E}[X_i]. \quad (12)$$

- (e) *Regla del producto independiente.* Si las variables aleatorias X_1, \dots, X_n tienen esperanza finita y son independientes, entonces el producto tiene esperanza finita y coincide con el producto de las esperanzas:

$$\mathbb{E} \left[\prod_{i=1}^n X_i \right] = \prod_{i=1}^n \mathbb{E}[X_i]. \quad (13)$$

Demostración. (a) es consecuencia inmediata de la Definición 1.1 porque $\mathbb{P}(X = 1) = 1$.
(b) es consecuencia del Teorema 1.11 y de que para todo $x \in \mathbb{R}$ vale que $F_{X_1}(x) \geq F_{X_2}(x)$.
(c) es consecuencia inmediata del Teorema 1.16. (d) es consecuencia inmediata del Teorema 1.19. (e) es consecuencia del Teorema 1.19 y de la factorización de la distribución conjunta como producto de las distribuciones marginales. \square

1.4. Dividir y conquistar

Teorema 1.21. Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad y sea $X : \Omega \rightarrow \mathbb{R}$ una variable aleatoria. Sea $A \subset \mathbb{R}$ un conjunto tal que $\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{A}$. Si $\mathbb{P}(X \in A) > 0$, entonces

$$\mathbb{E}[X|X \in A] = \frac{1}{\mathbb{P}(X \in A)} \mathbb{E}[X \mathbf{1}\{X \in A\}]. \quad (14)$$

Demostración. Para simplificar la exposición vamos a suponer que la variable aleatoria X es discreta. Por la Definición 1.1 tenemos que

$$\begin{aligned} \mathbb{E}[X|X \in A] &= \sum_{x \in X(\Omega)} x p_{X|X \in A}(x) = \sum_{x \in X(\Omega)} x \frac{\mathbb{P}(X = x)}{\mathbb{P}(X \in A)} \mathbf{1}\{x \in A\} \\ &= \frac{1}{\mathbb{P}(X \in A)} \sum_{x \in X(\Omega)} x \mathbf{1}\{x \in A\} \mathbb{P}(X = x) = \frac{1}{\mathbb{P}(X \in A)} \mathbb{E}[X \mathbf{1}\{X \in A\}]. \end{aligned}$$

La última igualdad es consecuencia del Teorema 1.16. \square

Ejemplo 1.22. Sea X el resultado del tiro de un dado equilibrado y sea $A = \{2, 4, 6\}$. De acuerdo con (14) la esperanza de $X|X \in A$ es

$$\mathbb{E}[X|X \in A] = \frac{1}{\mathbb{P}(X \in A)} \mathbb{E}[X \mathbf{1}\{X \in A\}] = \frac{1}{1/2} \left(\frac{2}{6} + \frac{4}{6} + \frac{6}{6} \right) = 4.$$

Resultado que por otra parte es intuitivamente evidente. \square

Teorema 1.23 (Fórmula de probabilidad total). Sea X una variable aleatoria. Si A_1, \dots, A_n es una partición medible de \mathbb{R} tal que $\mathbb{P}(X \in A_i) > 0$, $i = 1, \dots, n$. Entonces,

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X|X \in A_i] \mathbb{P}(X \in A_i). \quad (15)$$

Demostración. Descomponemos la variable X como una suma de variables (dependientes de la partición) $X = \sum_{i=1}^n X \mathbf{1}\{X \in A_i\}$. Como la esperanza es un operador lineal tenemos que

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X \mathbf{1}\{X \in A_i\}] = \sum_{i=1}^n \mathbb{E}[X|X \in A_i] \mathbb{P}(X \in A_i).$$

La última igualdad se obtiene de (14). \square

Nota Bene. Sea $g : \mathbb{R} \rightarrow \mathbb{R}$ una función tal que $g(X)$ es una variable aleatoria. Bajo las hipótesis del Teorema 1.23 también vale que

$$\mathbb{E}[g(X)] = \sum_{i=1}^n \mathbb{E}[g(X)|X \in A_i] \mathbb{P}(X \in A_i). \quad (16)$$

La fórmula (16) se puede extender sin ninguna dificultad al caso multidimensional. \square

Ejemplo 1.24 (Dividir y conquistar). Todas las mañanas Lucas llega a la estación del subte entre las 7:10 y las 7:30 (con distribución uniforme en el intervalo). El subte llega a la estación cada quince minutos comenzando a las 6:00. Calcular la media del tiempo que tiene que esperar Lucas hasta subirse al subte.

Sea X el horario en que Lucas llega a la estación del subte. El tiempo que tiene que esperar hasta subirse al subte se describe por

$$T = (7.15 - X)\mathbf{1}\{X \in [7:10, 7:15]\} + (7:30 - X)\mathbf{1}\{X \in (7:15, 7:30]\}.$$

Ahora bien, dado que $X \in [7:10, 7:15]$, la distribución de T es uniforme sobre el intervalo $[0, 5]$ minutos y dado que $X \in (7:15, 7:30]$ la distribución de T es uniforme sobre el intervalo $[0, 15]$ minutos. De acuerdo con (16)

$$\mathbb{E}[T] = \frac{5}{2} \left(\frac{5}{20} \right) + \frac{15}{2} \left(\frac{15}{20} \right) = 6.25.$$

□

2. Varianza

2.1. Definición

La esperanza de una variable aleatoria X , $\mathbb{E}[X]$, también se conoce como *la media* o el primer momento de X . La cantidad $\mathbb{E}[X^n]$, $n \geq 1$, se llama el *n-ésimo momento* de X . Si la esperanza $\mathbb{E}[X]$ es finita, la cantidad $\mathbb{E}[(X - \mathbb{E}[X])^n]$ se llama el *n-ésimo momento central*.

Después de la esperanza la siguiente cantidad en orden de importancia para resumir el comportamiento de una variable aleatoria X es su segundo momento central también llamado la *varianza de X* .

Definición 2.1 (Varianza). Sea X una variable aleatoria con esperanza finita. La *varianza* de X se define por

$$\mathbb{V}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]. \quad (17)$$

En otras palabras, la varianza de X es la esperanza de la variable aleatoria $(X - \mathbb{E}[X])^2$. Puesto que $(X - \mathbb{E}[X])^2$ sólo puede tomar valores no negativos, la varianza es no negativa.

La varianza de X es una de las formas más utilizadas para medir la dispersión de los valores de X respecto de su media. Otra medida de dispersión es el *desvío estándar* de X , que se define como la raíz cuadrada de la varianza y se denota $\sigma(X)$:

$$\sigma(X) := \sqrt{\mathbb{V}(X)}. \quad (18)$$

A diferencia de la varianza, el desvío estándar de una variable aleatoria es más fácil de interpretar porque tiene las mismas unidades de X .

Nota Bene: Grandes valores de $\mathbb{V}(X)$ significan grandes variaciones de los valores de X alrededor de la media. Al contrario, pequeños valores de $\mathbb{V}(X)$ implican una pronunciada concentración de la masa de la distribución de probabilidades en un entorno de la media. En el caso extremo, cuando la varianza es 0, la masa total de la distribución de probabilidades se concentra en la media. Estas afirmaciones pueden hacerse más precisas y serán desarrolladas en la sección 4.

2.2. Cálculo

Una manera “brutal” de calcular $\mathbb{V}(X)$ es calcular la función de distribución de la variable aleatoria $(X - \mathbb{E}[X])^2$ y usar la definición de esperanza. En lo que sigue mostraremos una manera más simple de realizar ese tipo cálculo.

Proposición 2.2 (Expresión de la varianza en términos de los momentos). Sea X una variable aleatoria con primer y segundo momentos finitos, entonces

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \quad (19)$$

En palabras, la varianza es la diferencia entre el segundo momento y el cuadrado del primer momento.

Demostración. Desarrollar el cuadrado $(X - \mathbb{E}[X])^2$ y usar las propiedades de la esperanza. Poniendo $(X - \mathbb{E}[X])^2 = X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2$ se obtiene

$$\mathbb{V}(X) = \mathbb{E}[X^2] - 2X\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

□

Ejemplo 2.3 (Varianza de la función indicadora). Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad. Para cualquier evento $A \in \mathcal{A}$ vale que

$$\mathbb{V}(\mathbf{1}\{\omega \in A\}) = \mathbb{E}[\mathbf{1}\{\omega \in A\}^2] - \mathbb{E}[\mathbf{1}\{\omega \in A\}]^2 = \mathbb{P}(A) - \mathbb{P}(A)^2 = \mathbb{P}(A)(1 - \mathbb{P}(A)). \quad (20)$$

□

Ejemplo 2.4 (Dado equilibrado). Sea X el resultado del lanzamiento de un dado equilibrado. Por el Ejemplo 1.3 sabemos que $\mathbb{E}[X] = 7/2$. Por otra parte

$$\mathbb{E}[X^2] = \sum_{x=1}^6 x^2 \mathbb{P}(X = x) = \frac{1}{6} \sum_{x=1}^6 x^2 = \frac{1+4+9+16+25+36}{6} = \frac{91}{6}.$$

Por lo tanto, de acuerdo con la Proposición 2.2, la varianza de X es

$$\mathbb{V}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{32}{12} = \frac{8}{3}.$$

□

Ejemplo 2.5 (Fiabilidad). Sea T el tiempo de espera hasta que ocurre la primer falla en un sistema electrónico con función intensidad de fallas de la forma $\lambda(t) = 2t\mathbf{1}\{t > 0\}$. Por el Ejemplo 1.7 sabemos que $\mathbb{E}[T] = \sqrt{\pi}/2$. Por otra parte,

$$\mathbb{E}[T^2] = \int_{-\infty}^{\infty} t^2 f(t) dt = \int_0^{\infty} t^2 2t \exp(-t^2) dt = \int_0^{\infty} x e^{-x} dx = 1.$$

La tercera igualdad se obtiene mediante el cambio de variables $t^2 = x$ y la cuarta se deduce usando la fórmula de integración por partes aplicada a $u = x$ y $v' = e^{-x}$.

Por lo tanto, de acuerdo con la Proposición 2.2, la varianza de T es

$$\mathbb{V}(T) = 1 - \left(\frac{\sqrt{\pi}}{2}\right)^2 = 1 - \frac{\pi}{4}.$$

□

2.3. Propiedades

Proposición 2.6. Para todo $a, b \in \mathbb{R}$

$$\mathbb{V}(aX + b) = a^2\mathbb{V}(X). \quad (21)$$

Demostración. Por definición,

$$\mathbb{V}(aX + b) = \mathbb{E}[(aX + b - \mathbb{E}[aX + b])^2] = \mathbb{E}[a^2(X - \mathbb{E}[X])^2] = a^2\mathbb{V}(X).$$

Para obtener la segunda igualdad usamos que $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$. □

Error cuadrático medio. Una manera de “representar” la variable aleatoria X mediante un valor fijo $c \in \mathbb{R}$ es hallar el valor c que minimice el llamado *error cuadrático medio*, $\mathbb{E}[(X - c)^2]$.

Teorema 2.7 (Pitágoras). Sea X una variable aleatoria con esperanza y varianza finitas. Para toda constante $c \in \mathbb{R}$ vale que

$$\mathbb{E}[(X - c)^2] = \mathbb{V}(X)^2 + (\mathbb{E}[X] - c)^2.$$

En particular, el valor de c que minimiza el error cuadrático medio es la esperanza de X , $\mathbb{E}[X]$.

Demostración. Escribiendo $X - c$ en la forma $X - \mathbb{E}[X] + \mathbb{E}[X] - c$ y desarrollando cuadrados se obtiene $(X - c)^2 = (X - \mathbb{E}[X])^2 + (\mathbb{E}[X] - c)^2 + 2(X - \mathbb{E}[X])(\mathbb{E}[X] - c)$. El resultado se obtiene tomando esperanza en ambos lados de la igualdad y observando que $\mathbb{E}[X - \mathbb{E}[X]] = 0$. □

3. Covarianza

3.1. Definición

Definición 3.1 (Covarianza). Sean X e Y dos variables aleatorias de varianzas finitas definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. La *covarianza* de X e Y se define por

$$Cov(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]. \quad (22)$$

3.2. Cálculo

Proposición 3.2. Sean X e Y dos variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Si los segundos momentos de las variables aleatorias X e Y son finitos, se tiene que

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \quad (23)$$

Demuestra. La esperanza del producto $E[XY]$ es finita porque las esperanzas $\mathbb{E}[X^2]$ y $\mathbb{E}[Y^2]$ son finitas y vale que $|xy| \leq \frac{1}{2}(x^2 + y^2)$. Usando la propiedad distributiva del producto y la linealidad de la esperanza tenemos que

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] &= \mathbb{E}[XY - \mathbb{E}[Y]X - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[Y]\mathbb{E}[X] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

□

Ejemplo 3.3. Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad y sean $A \in \mathcal{A}$ y $B \in \mathcal{A}$ dos eventos de probabilidad positiva. Consideremos las variables aleatorias $X = \mathbf{1}\{\omega \in A\}$ e $Y = \mathbf{1}\{\omega \in B\}$. Entonces,

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{P}(XY = 1) - \mathbb{P}(X = 1)\mathbb{P}(Y = 1) \\ &= \mathbb{P}(X = 1, Y = 1) - \mathbb{P}(X = 1)\mathbb{P}(Y = 1).\end{aligned}$$

La segunda y la tercera igualdad se obtienen de (2) observando que XY es una variable a valores 0 o 1 que vale 1 si y solo si X e Y son ambas 1.

Notamos que

$$\begin{aligned}\text{Cov}(X, Y) > 0 &\iff \mathbb{P}(X = 1, Y = 1) > \mathbb{P}(X = 1)\mathbb{P}(Y = 1) \\ &\iff \frac{\mathbb{P}(X = 1, Y = 1)}{\mathbb{P}(X = 1)} > \mathbb{P}(Y = 1) \\ &\iff \mathbb{P}(Y = 1|X = 1) > \mathbb{P}(Y = 1).\end{aligned}$$

En palabras, la covarianza de X e Y es positiva si y solamente si la condición $X = 1$ aumenta la probabilidad de que $Y = 1$. □

Ejemplo 3.4. En una urna hay 6 bolas rojas y 4 bolas negras. Se extraen 2 bolas al azar sin reposición. Consideramos los eventos

$$A_i = \{\text{sale una bola roja en la } i\text{-ésima extracción}\}, \quad i = 1, 2,$$

y definimos las variables aleatorias X_1 y X_2 como las funciones indicadoras de los eventos A_1 y A_2 respectivamente. De acuerdo con el Ejemplo anterior es intuitivamente claro que $\text{Cov}(X_1, X_2) < 0$. (*¿Por qué?*)

$$\begin{aligned}\text{Cov}(X_1, X_2) &= \mathbb{P}(X_1 = 1, X_2 = 1) - \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1) = \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1)\mathbb{P}(A_2) \\ &= \frac{6}{10} \times \frac{5}{9} - \frac{6}{10} \left(\frac{5}{9} \times \frac{6}{10} + \frac{6}{9} \times \frac{4}{10} \right) = -\frac{2}{75} = -0.02666....\end{aligned}$$

□

Nota Bene. Se puede mostrar que $\text{Cov}(X, Y) > 0$ es una indicación de que Y tiende a crecer cuando X lo hace, mientras que $\text{Cov}(X, Y) < 0$ es una indicación de que Y decrece cuando X crece. □

3.3. Propiedades

Lema 3.5 (Propiedades). Para variables aleatorias X, Y, Z y constantes a , valen las siguientes propiedades

1. $\text{Cov}(X, X) = \mathbb{V}(X)$,
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$,
3. $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$,
4. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$.

Demostración. *Ejercicio.*

□

Sobre la esperanza del producto. Si se conoce la covarianza y la esperanza de las marginales, la identidad (23) puede ser útil para calcular la esperanza del producto:

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + \text{Cov}(X, Y).$$

Nota Bene. Si X e Y son independientes, $\text{Cov}(X, Y) = 0$ porque $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Pero la recíproca no es cierta.

Ejemplo 3.6 (Dos bolas en dos urnas). El experimento aleatorio consiste en ubicar dos bolas distinguibles en dos urnas. Sean N la cantidad de urnas ocupadas y X_i la cantidad de bolas en la urna i . El espacio muestral se puede representar de la siguiente manera $\Omega = \{(1, 1); (1, 2); (2, 1); (2, 2)\}$. La función de probabilidad conjunta de N y X_1 se muestra en el Cuadro 1

$N \setminus X_1$	0	1	2	p_N
1	1/4	0	1/4	1/2
2	0	1/2	0	1/2
p_{X_1}	1/4	1/2	1/4	

Cuadro 1: Función de probabilidad conjunta de (N, X_1) .

Para calcular la esperanza del producto NX_1 usamos el Teorema 1.19

$$\begin{aligned} \mathbb{E}[NX_1] &= 1 \cdot 1 \cdot p_{N, X_1}(1, 1) + 1 \cdot 2 \cdot p_{N, X_1}(1, 2) + 2 \cdot 1 \cdot p_{N, X_1}(2, 1) + 2 \cdot 2 \cdot p_{N, X_1}(2, 2) \\ &= 1 \cdot 0 + 2 \cdot 1/4 + 2 \cdot 1/2 + 4 \cdot 0 = 3/2. \end{aligned}$$

Es fácil ver que $\mathbb{E}[N] = 3/2$ y $\mathbb{E}[X_1] = 1$. Por lo tanto, $\text{Cov}(N, X_1) = 0$. Sin embargo, las variables N y X_1 no son independientes. □

3.4. Varianza de sumas

Usando las propiedades de la covarianza enunciadas en Lema 3.5 se puede demostrar que

$$\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j) \quad (24)$$

En particular, se obtiene que

$$\mathbb{V} \left(\sum_{i=1}^n X_i \right) = Cov \left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j \right) = \sum_{i=1}^n \mathbb{V}(X_i) + 2 \sum_{i=1}^n \sum_{j < i} Cov(X_i, Y_j). \quad (25)$$

Finalmente, si las variables son independientes

$$\mathbb{V} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \mathbb{V}(X_i). \quad (26)$$

4. Algunas desigualdades

4.1. Cauchy-Schwartz

Teorema 4.1 (Cauchy-Schwartz).

$$\mathbb{E}[|XY|] \leq (\mathbb{E}[X^2]\mathbb{E}[Y^2])^{1/2} \quad (27)$$

Demostración. Observar que para todo $t \in \mathbb{R}$:

$$0 \leq \mathbb{E}[(t|X| + |Y|)^2] = t^2\mathbb{E}[X^2] + 2t\mathbb{E}[|XY|] + \mathbb{E}[Y^2].$$

Como la función cuadrática en t que aparece en el lado derecho de la igualdad tiene a lo sumo una raíz real se deduce que

$$4\mathbb{E}[|XY|]^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0.$$

Por lo tanto,

$$\mathbb{E}[|XY|]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

□

Corolario 4.2. Sea X una variable aleatoria tal que $\mathbb{E}[X^2] < \infty$. Si $a < \mathbb{E}[X]$, entonces

$$\mathbb{P}(X > a) \geq \frac{(\mathbb{E}[X] - a)^2}{\mathbb{E}[X^2]}.$$

Demostración. De la desigualdad $X\mathbf{1}\{X > a\} \leq |X\mathbf{1}\{X > a\}|$ y de la propiedad de monotonía de la esperanza se deduce que

$$\mathbb{E}[X\mathbf{1}\{X > a\}] \leq E[|X\mathbf{1}\{X > a\}|]. \quad (28)$$

Aplicando la desigualdad de Cauchy-Schwartz a $|X\mathbf{1}\{X > a\}|$ se obtiene que

$$\mathbb{E}[|X\mathbf{1}\{X > a\}|] \leq (\mathbb{E}[X^2]\mathbb{E}[\mathbf{1}\{X > a\}^2])^{1/2} = (\mathbb{E}[X^2]\mathbb{P}(X > a))^{1/2} \quad (29)$$

Observando que $X = X\mathbf{1}\{X > a\} + X\mathbf{1}\{X \leq a\}$ y que $X\mathbf{1}\{X \leq a\} \leq a$ se deduce que

$$\mathbb{E}[X] = \mathbb{E}[X\mathbf{1}\{X > a\}] + \mathbb{E}[X\mathbf{1}\{X \leq a\}] \leq \mathbb{E}[X\mathbf{1}\{X > a\}] + a$$

y en consecuencia,

$$\mathbb{E}[X] - a \leq \mathbb{E}[X \mathbf{1}\{X > a\}]. \quad (30)$$

Combinando las desigualdades (30), (28) y (29) se obtiene que

$$\mathbb{E}[X] - a \leq (\mathbb{E}[X^2]\mathbb{P}(X > a))^{1/2}$$

y como $\mathbb{E}[X] - a > 0$, elevando al cuadrado, se concluye que

$$(\mathbb{E}[X] - a)^2 \leq \mathbb{E}[X^2]\mathbb{P}(X > a).$$

El resultado se obtiene despejando. \square

4.2. Chebyshev

Teorema 4.3 (Desigualdad de Chebyshev). *Sea $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ tal que $\varphi \geq 0$ y $A \in \mathcal{B}(\mathbb{R})$. Sea $i_A := \inf\{\varphi(x) : x \in A\}$. Entonces,*

$$i_A \mathbb{P}(X \in A) \leq \mathbb{E}[\varphi(X)] \quad (31)$$

Demostración. La definición de i_A y el hecho de que $\varphi \geq 0$ implican que

$$i_A \mathbf{1}\{X \in A\} \leq \varphi(X) \mathbf{1}\{X \in A\} \leq \varphi(X)$$

El resultado se obtiene tomando esperanza. \square

En lo que sigue enunciaremos algunos corolarios que se obtienen como casos particulares del Teorema 4.3.

Corolario 4.4 (Desigualdad de Markov). *Sea X una variable aleatoria a valores no negativos. Para cada $a > 0$ vale que*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \quad (32)$$

Demostración. Aplicar la desigualdad de Chebyshev usando la función $\varphi(x) = x$ restringida a la semi-recta no negativa $[0, \infty)$ y el conjunto $A = [a, \infty)$ para obtener

$$a \mathbb{P}(X \geq a) \leq \mathbb{E}[\varphi(X)] = \mathbb{E}[X].$$

y despejar. \square

Corolario 4.5. *Sea $a > 0$. Vale que*

$$\mathbb{P}(X > a) \leq \frac{1}{a^2} \mathbb{E}[X^2]. \quad (33)$$

Demostración. Aplicar la desigualdad de Chebyshev usando la función $\varphi(x) = x^2$ y el conjunto $A = (a, \infty)$ para obtener

$$a^2 \mathbb{P}(X > a) \leq \mathbb{E}[X^2]$$

y despejar. \square

Corolario 4.6 (Pequeña desigualdad de Chebyshev). *Sea X una variable aleatoria de varianza finita. Para cada $a > 0$ vale que*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbb{V}(X)}{a^2}. \quad (34)$$

Demostración. Debido a que $(X - \mathbb{E}[X])^2$ es una variable aleatoria no negativa podemos aplicar la desigualdad de Markov (poniendo a^2 en lugar de a) y obtenemos

$$\mathbb{P}((X - \mathbb{E}[X])^2 \geq a^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2} = \frac{\mathbb{V}(X)}{a^2}.$$

La desigualdad $(X - \mathbb{E}[X])^2 \geq a^2$ es equivalente a la desigualdad $|X - \mathbb{E}[X]| \geq a$. Por lo tanto,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbb{V}(X)}{a^2}.$$

Lo que concluye la demostración. \square

Nota Bene. *Grosso modo* la pequeña desigualdad de Chebyshev establece que si la varianza es pequeña, los grandes desvíos respecto de la media son improbables.

Corolario 4.7. Sea X una variable aleatoria con varianza finita, entonces para cada $\alpha > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \alpha\sigma(X)) \leq \frac{1}{\alpha^2}. \quad (35)$$

El resultado se obtiene poniendo $a = \alpha\sigma(X)$ en la pequeña desigualdad de Chebyshev. \square

Ejemplo 4.8. La cantidad X de artículos producidos por un fábrica durante una semana es una variable aleatoria de media 500.

(a) ¿Qué puede decirse sobre la probabilidad de que la producción semanal supere los 1000 artículos? Por la desigualdad de Markov,

$$\mathbb{P}(X \geq 1000) \leq \frac{\mathbb{E}[X]}{1000} = \frac{500}{1000} = \frac{1}{2}.$$

(b) Si la varianza de la producción semanal es conocida e igual a 100, ¿qué puede decirse sobre la probabilidad de que la producción semanal se encuentre entre 400 y 600 artículos? Por la desigualdad de Chebyshev,

$$\mathbb{P}(|X - 500| \geq 100) \leq \frac{\sigma^2}{(100)^2} = \frac{1}{100}.$$

Por lo tanto, $\mathbb{P}(|X - 500| < 100) \geq 1 - \frac{1}{100} = \frac{99}{100}$, la probabilidad de que la producción semanal se encuentre entre 400 y 600 artículos es al menos 0.99. \square

El que mucho abarca poco aprieta. Las desigualdades de Markov y Chebyshev son importantes porque nos permiten deducir cotas sobre las probabilidades cuando solo se conocen la media o la media y la varianza de la distribución de probabilidades. Sin embargo, debe tenerse en cuenta que las desigualdades de Markov y de Chebyshev producen cotas universales que no dependen de las distribuciones de las variables aleatorias (dependen pura y exclusivamente de los valores de la esperanza y de la varianza). Por este motivo su comportamiento será bastante heterogéneo: en algunos casos producirán cotas extremadamente finas, pero en otros casos solamente cotas groseras. \square

5. La ley débil de los grandes números

Teorema 5.1 (Ley débil de los grandes números). Sea X_1, X_2, \dots una sucesión de variables aleatorias independientes idénticamente distribuidas, tales que $\mathbb{V}(X_1) < \infty$. Sea $S_n, n \geq 1$, la sucesión de las sumas parciales definida por $S_n := \sum_{i=1}^n X_i$. Entonces, para cualquier $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\left| \frac{S_n}{n} - \mathbb{E}[X_1] \right| > \epsilon \right) = 0.$$

Demostración. Se obtiene aplicando la desigualdad de Chebyshev a la variable aleatoria S_n/n . Usando que la esperanza es un operador lineal se obtiene que

$$\mathbb{E}[S_n/n] = \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mathbb{E}[X_1].$$

Como las variables X_1, X_2, \dots son independientes tenemos que

$$\mathbb{V}(S_n/n) = \frac{1}{n^2} \mathbb{V} \left(\sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}(X_i) = \frac{\mathbb{V}(X_1)}{n}.$$

Entonces, por la desigualdad de Chebyshev, obtenemos la siguiente estimación

$$\mathbb{P} \left(\left| \frac{S_n}{n} - \mathbb{E}[X_1] \right| > \epsilon \right) \leq \frac{\mathbb{V}(X_1)}{n\epsilon^2}. \quad (36)$$

Como $\mathbb{V}(X_1) < \infty$ el lado derecho de la última desigualdad tiende a 0 cuando $n \rightarrow \infty$. □

Nota Bene. La ley débil de los grandes números establecida en el Teorema 5.1 sirve como base para la noción intuitiva de probabilidad como medida de las frecuencias relativas. La proposición “en una larga serie de ensayos idénticos la frecuencia relativa del evento A se approxima a su probabilidad $\mathbb{P}(A)$ ” se puede hacer teóricamente más precisa de la siguiente manera: el resultado de cada ensayo se representa por una variable aleatoria (independiente de las demás) que vale 1 cuando se obtiene el evento A y vale cero en caso contrario. La expresión “una larga serie de ensayos” adopta la forma de una sucesión X_1, X_2, \dots de variables aleatorias independientes cada una con la misma distribución que la indicadora del evento A . Notar que $X_i = 1$ significa que “en el i -ésimo ensayo ocurrió el evento A ” y la suma parcial $S_n = \sum_{i=1}^n X_i$ representa la “frecuencia del evento A ” en los primeros n ensayos. Puesto que $\mathbb{E}[X_1] = \mathbb{P}(A)$ y $\mathbb{V}(X_1) = \mathbb{P}(A)(1 - \mathbb{P}(A))$ la estimación (36) adopta la forma

$$\mathbb{P} \left(\left| \frac{S_n}{n} - \mathbb{P}(A) \right| > \epsilon \right) \leq \frac{\mathbb{P}(A)(1 - \mathbb{P}(A))}{n\epsilon^2}. \quad (37)$$

Por lo tanto, *la probabilidad de que la frecuencia relativa del evento A se desvíe de su probabilidad $\mathbb{P}(A)$ en más de una cantidad prefijada ϵ , puede hacerse todo lo chica que se quiera, siempre que la cantidad de ensayos n sea suficientemente grande.*

Ejemplo 5.2 (Encuesta electoral). Se quiere estimar la proporción del electorado que prefiere votar a un cierto candidato. Cuál debe ser el tamaño muestral para garantizar un determinado *error* entre la proporción poblacional, p , y la proporción muestral S_n/n ?

Antes de resolver este problema, debemos reflexionar sobre la definición de *error*. Habitualmente, cuando se habla de error, se trata de un número real que expresa la (in)capacidad de una cierta cantidad de representar a otra. En los problemas de estimación estadística, debido a que una de las cantidades es una variable aleatoria y la otra no lo es, no es posible interpretar de un modo tan sencillo el significado de la palabra *error*.

Toda medida muestral tiene asociada una incerteza (o un riesgo) expresada por un modelo probabilístico. En este problema consideramos que el voto de cada elector se comporta como una variable aleatoria X tal que $\mathbb{P}(X = 1) = p$ y $\mathbb{P}(X = 0) = 1 - p$, donde $X = 1$ significa que el elector vota por el candidato considerado. Por lo tanto, cuando se habla de que queremos encontrar un tamaño muestral suficiente para un determinado error máximo, por ejemplo 0.02, tenemos que hacerlo con una medida de certeza asociada. Matemáticamente, queremos encontrar n tal que $\mathbb{P} \left(\left| \frac{S_n}{n} - p \right| \leq 0.02 \right) \geq 0.9999$ o, equivalentemente, queremos encontrar n tal que

$$\mathbb{P} \left(\left| \frac{S_n}{n} - p \right| > 0.02 \right) \leq 0.0001.$$

Usando la estimación (37) se deduce que

$$\mathbb{P} \left(\left| \frac{S_n}{n} - p \right| > 0.02 \right) \leq \frac{p(1-p)}{n(0.02)^2}.$$

El numerador de la fracción que aparece en el lado derecho de la estimación depende de p y el valor de p es desconocido. Sin embargo, sabemos que $p(1-p)$ es una parábola convexa con raíces en $p = 0$ y $p = 1$ y por lo tanto su máximo ocurre cuando $p = 1/2$, esto es $p(1-p) \leq 1/4$. En la peor hipótesis tenemos:

$$\mathbb{P} \left(\left| \frac{S_n}{n} - p \right| > 0.02 \right) \leq \frac{1}{4n(0.02)^2}.$$

Como máximo estamos dispuestos a correr un riesgo de 0.0001 y en el peor caso tenemos acotada la máxima incerteza por $(4n(0.02)^2)^{-1}$. El problema se reduce a resolver la desigualdad $(4n(0.02)^2)^{-1} \leq 0.0001$. Por lo tanto,

$$n \geq ((0.0001)^{-1}) / (4(0.02)^2) = 6250000.$$

Una cifra absurdamente grande!! Más adelante, mostraremos que existen métodos más sofisticados que permiten disminuir el tamaño de la muestra. \square

6. Distribuciones particulares

Para facilitar referencias posteriores presentaremos tablas de esperanzas y varianzas de algunas distribuciones importantes de uso frecuente y describiremos el método para obtenerlas.

Discretas

No.	Nombre	Probabilidad	Soporte	Esperanza	Varianza
1.	Uniforme	$\frac{1}{b-a+1}$	$a \leq x \leq b$	$(a+b)/2$	$(b-a)(b-a-2)/12$
2.	Bernoulli	$p^x(1-p)^{1-x}$	$x \in \{0, 1\}$	p	$p(1-p)$
3.	Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	$0 \leq x \leq n$	np	$np(1-p)$
4.	Geométrica	$(1-p)^{x-1}p$	$x \in \mathbb{N}$	$1/p$	$(1-p)/p^2$
5.	Poisson	$\frac{\lambda^x}{x!} e^{-\lambda}$	$x \in \mathbb{N}_0$	λ	λ

Cuadro 2: Esperanza y varianza de algunas distribuciones discretas de uso frecuente.

Continuas

No.	Nombre	Densidad	Soporte	Esperanza	Varianza
1.	Uniforme	$\frac{1}{b-a}$	$x \in [a, b]$	$(a+b)/2$	$(b-a)^2/12$
2.	Exponencial	$\lambda e^{-\lambda x}$	$x > 0$	$1/\lambda$	$1/\lambda^2$
3.	Gamma	$\frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}$	$x > 0$	ν/λ	ν/λ^2
4.	Beta	$\frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1} (1-x)^{\nu_2-1}$	$x \in (0, 1)$	$\frac{\nu_1}{\nu_1+\nu_2}$	$\frac{\nu_1\nu_2}{(\nu_1+\nu_2)^2(\nu_1+\nu_2+1)}$
5.	Normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$	$x \in \mathbb{R}$	μ	σ^2

Cuadro 3: Esperanza y varianza de algunas distribuciones continuas de uso frecuente.

Cuentas con variables discretas

1. Distribución uniforme discreta.

Sean a y b dos números enteros tales que $a < b$. Se dice que la variable aleatoria X tiene distribución uniforme sobre el “intervalo” de números enteros $[a, b] := \{a, a+1, \dots, b\}$, y se denota $X \sim \mathcal{U}[a, b]$, si X es discreta y tal que

$$\mathbb{P}(X = x) = \frac{1}{b-a+1} \mathbf{1}\{x \in \{a, a+1, \dots, b\}\}.$$

Notando que la distribución de X coincide con la de la variable $X^* + a - 1$, donde X^* está uniformemente distribuida sobre $\{1, \dots, b-a+1\}$, resulta que

$$\mathbb{E}[X] = \mathbb{E}[X^*] + a - 1 = \frac{1 + (b-a+1)}{2} + a - 1 = \frac{a+b}{2}.$$

Para calcular la varianza de X , consideramos primero el caso más simple donde $a = 1$ y $b = n$. Por inducción en n se puede ver que

$$\mathbb{E}[X^2] = \frac{1}{n} \sum_{k=1}^n k^2 = \frac{(n+1)(2n+1)}{6}.$$

La varianza puede obtenerse en términos de los momentos de orden 1 y 2:

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{(n+1)[2(2n+1) - 3(n+1)]}{12} = \frac{n^2 - 1}{12}. \end{aligned}$$

Para el caso general, notamos que la variable aleatoria uniformemente distribuida sobre $[a, b]$ tiene la misma varianza que la variable aleatoria uniformemente distribuida sobre $[1, b-a+1]$, puesto que esas dos variables difieren en la constante $a-1$. Por lo tanto, la varianza buscada se obtiene de la fórmula anterior sustituyendo $n = b-a+1$

$$\mathbb{V}(X) = \frac{(b-a+1)^2 - 1}{12} = \frac{(b-a)(b-a+2)}{12}.$$

□

2. Distribución Bernoulli.

Sea $p \in (0, 1)$. Se dice que la variable aleatoria X tiene distribución *Bernoulli de parámetro p* , y se denota $X \sim \text{Bernoulli}(p)$, si X es discreta y tal que

$$\mathbb{P}(X = x) = p^x(1-p)^{1-x}, \text{ donde } x = 0, 1.$$

Por definición,

$$\mathbb{E}[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Por otra parte,

$$\mathbb{E}[X^2] = 0^2 \cdot \mathbb{P}(X = 0) + 1^2 \cdot \mathbb{P}(X = 1) = p.$$

Por lo tanto,

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p).$$

□

3. Distribución Binomial.

Sean $p \in (0, 1)$ y $n \in \mathbb{N}$. Se dice que la variable aleatoria X tiene distribución *Binomial de parámetros n y p* , y se denota $X \sim \text{Binomial}(n, p)$, si X es discreta y tal que

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ donde } x = 0, 1, \dots, n.$$

Por definición,

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=0}^n x \mathbb{P}(X = x) = \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^n \frac{x n!}{(n-x)! x!} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^n \frac{n!}{(n-x)!(x-1)!} p^x (1-p)^{n-x} = np \sum_{x=1}^n \frac{(n-1)!}{(n-x)!(x-1)!} p^{x-1} (1-p)^{n-x} \\ &= np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y} = np(p + (1-p))^{n-1} = np.\end{aligned}$$

Análogamente se puede ver que

$$\mathbb{E}[X^2] = np((n-1)p + 1).$$

Por lo tanto,

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = np((n-1)p + 1) - (np)^2 \\ &= np((n-1)p + 1 - np) = np(1-p).\end{aligned}$$

□

4. Distribución Geométrica.

Sea $p \in (0, 1)$. Se dice que la variable aleatoria X tiene distribución *Geométrica de parámetro p* , y se denota $X \sim \text{Geométrica}(p)$, si X es discreta y tal que

$$\mathbb{P}(X = x) = (1-p)^{x-1} p \mathbf{1}\{x \in \mathbb{N}\}.$$

Por definición,

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x \mathbb{P}(X = x) = \sum_{x=1}^{\infty} x (1-p)^{x-1} p = p \sum_{x=1}^{\infty} x (1-p)^{x-1}.$$

La serie se calcula observando que $x(1-p)^{x-1} = -\frac{d}{dp}(1-p)^x$ y recordando que las series de potencias se pueden derivar término a término:

$$\sum_{x=1}^{\infty} x (1-p)^{x-1} = -\frac{d}{dp} \sum_{x=1}^{\infty} (1-p)^x = -\frac{d}{dp} (p^{-1} - 1) = p^{-2}.$$

Por lo tanto, $\mathbb{E}[X] = p \cdot p^{-2} = 1/p$.

Para calcular $\mathbb{V}(X)$ usaremos la misma técnica: derivamos dos veces ambos lados de la igualdad $\sum_{x=1}^{\infty} (1-p)^{x-1} = p^{-1}$ y obtenemos

$$\begin{aligned} 2p^{-3} &= \frac{d^2}{dp^2} p^{-1} = \frac{d^2}{dp^2} \sum_{x=1}^{\infty} (1-p)^{x-1} = \sum_{x=1}^{\infty} (x-1)(x-2)(1-p)^{x-3} \\ &= \sum_{x=1}^{\infty} (x+1)x(1-p)^{x-1} = \sum_{x=1}^{\infty} x^2(1-p)^{x-1} + \sum_{x=1}^{\infty} x(1-p)^{x-1}. \end{aligned}$$

Multiplicando por p los miembros de las igualdades obtenemos, $2p^{-2} = \mathbb{E}[X^2] + \mathbb{E}[X] = \mathbb{E}[X^2] + p^{-1}$. En consecuencia, $\mathbb{E}[X^2] = 2p^{-2} - p^{-1}$. Por lo tanto,

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 2p^{-2} - p^{-1} - p^{-2} = p^{-2} - p^{-1} = p^{-2}(1-p).$$

□

5. Distribución de Poisson.

Sea $\lambda > 0$. Se dice que la variable aleatoria X tiene distribución de *Poisson de intensidad* λ , y se denota $X \sim \text{Poisson}(\lambda)$, si X es discreta y tal que

$$\mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \mathbf{1}\{x \in \mathbb{N}_0\}.$$

Por definición,

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x=0}^{\infty} x \mathbb{P}(X = x) = \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} x \frac{\lambda^{x-1}}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda. \end{aligned}$$

Derivando término a término, se puede ver que

$$\begin{aligned} \mathbb{E}[X^2] &= \sum_{x=0}^{\infty} x^2 \mathbb{P}(X = x) = \sum_{x=0}^{\infty} x^2 \frac{\lambda^x}{x!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} x^2 \frac{\lambda^{x-1}}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{x \lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda} \frac{d}{d\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} = \lambda e^{-\lambda} \frac{d}{d\lambda} (\lambda e^{\lambda}) = \lambda e^{-\lambda} (e^{\lambda} + \lambda e^{\lambda}) = \lambda + \lambda^2. \end{aligned}$$

Por lo tanto,

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda + \lambda^2 - \lambda^2 = \lambda.$$

□

Cuentas con variables continuas

1. Distribución uniforme.

Sean $a < b$. Se dice que la variable aleatoria X tiene distribución *uniforme sobre el intervalo* $[a, b]$, y se denota $X \sim \mathcal{U}[a, b]$, si X es absolutamente continua con densidad de probabilidades

$$f(x) = \frac{1}{b-a} \mathbf{1}\{x \in [a, b]\}.$$

Por definición,

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} x \frac{1}{b-a} \mathbf{1}\{x \in [a, b]\} dx = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \left(\frac{b^2 - a^2}{2} \right) \\ &= \frac{a+b}{2}.\end{aligned}$$

Por otra parte,

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{1}{b-a} \int_a^b x^2 dx = \frac{1}{b-a} \left(\frac{b^3 - a^3}{3} \right) = \frac{a^2 + ab + b^2}{3}.$$

Finalmente,

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2} \right)^2 = \frac{a^2 - 2ab + b^2}{12} = \frac{(b-a)^2}{12}.$$

□

2. Distribución exponencial.

Sea $\lambda > 0$. Se dice que la variable aleatoria X tiene distribución *exponencial de intensidad* λ , y se denota $X \sim \text{Exp}(\lambda)$, si X es absolutamente continua con función densidad de probabilidades

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}\{x \geq 0\}.$$

El cálculo de $\mathbb{E}[X]$ y $\mathbb{V}(X)$ se reduce al caso $X \sim \text{Exp}(1)$. Basta observar que $Y \sim \text{Exp}(\lambda)$ si y solo si $Y = \lambda^{-1}X$, donde $X \sim \text{Exp}(1)$ y usar las identidades $\mathbb{E}[\lambda^{-1}X] = \lambda^{-1}\mathbb{E}[X]$ y $\mathbb{V}(\lambda^{-1}X) = \lambda^{-2}\mathbb{V}(X)$. En lo que sigue suponemos que $X \sim \text{Exp}(1)$.

Integrando por partes se obtiene,

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} xe^{-x} \mathbf{1}\{x \geq 0\} = \int_0^{\infty} \lambda xe^{-x} dx = -xe^{-x} \Big|_0^{\infty} + \int_0^{\infty} e^{-x} dx \\ &= 1.\end{aligned}$$

Por otra parte,

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{\infty} x^2 e^{-x} dx = -x^2 e^{-x} \Big|_0^{\infty} + \int_0^{\infty} 2xe^{-x} dx = 2.$$

Por lo tanto, $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 2 - 1 = 1$. □

3. Distribución gamma.

La *función gamma* se define por

$$\Gamma(t) := \int_0^{\infty} x^{t-1} e^{-x} dx \quad t > 0.$$

Integrando por partes puede verse que $\Gamma(t) = (t-1)\Gamma(t-1)$ para todo $t > 0$. De aquí se deduce que la función gamma interpola a los números factoriales en el sentido de que

$$\Gamma(n+1) = n! \quad \text{para } n = 0, 1, \dots$$

Sean $\lambda > 0$ y $\nu > 0$. Se dice que la variable aleatoria X tiene distribución *gamma de parámetros* ν, λ , y se denota $X \sim \Gamma(\nu, \lambda)$, si X es absolutamente continua con función densidad de probabilidades

$$f(x) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x} \mathbf{1}\{x > 0\}.$$

El cálculo de $\mathbb{E}[X]$ y $\mathbb{V}(X)$ se reduce al caso $X \sim \Gamma(\nu, 1)$. Para ello, basta observar que $Y \sim \Gamma(\nu, \lambda)$ si y solo si $Y = \lambda^{-1}X$, donde $X \sim \Gamma(\nu, 1)$ y usar las identidades $\mathbb{E}[\lambda^{-1}X] = \lambda^{-1}\mathbb{E}[X]$ y $\mathbb{V}(\lambda^{-1}X) = \lambda^{-2}\mathbb{V}(X)$. En lo que sigue suponemos que $X \sim \Gamma(\nu, 1)$

$$\mathbb{E}[X] = \int_0^\infty x f(x) dx = \int_0^\infty \frac{1}{\Gamma(\nu)} x^\nu e^{-x} dx = \frac{1}{\Gamma(\nu)} \Gamma(\nu + 1) = \nu.$$

Del mismo modo se puede ver que $\mathbb{E}[X^2] = (\nu + 1)\nu = \nu^2 + \nu$. Por lo tanto, $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \nu$. \square

4. Distribución beta

Sean $\nu_1 > 0$ y $\nu_2 > 0$. Se dice que la variable aleatoria X tiene distribución *beta de parámetros* ν_1, ν_2 , y se denota $X \sim \beta(\nu_1, \nu_2)$, si X es absolutamente continua con función densidad de probabilidades

$$f(x) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1} (1-x)^{\nu_2-1} \mathbf{1}\{x \in (0, 1)\}.$$

Por definición,

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^\infty x f(x) dx = \int_{-\infty}^\infty x \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1} (1-x)^{\nu_2-1} \mathbf{1}\{x \in (0, 1)\} dx \\ &= \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \int_0^1 x^{\nu_1} (1-x)^{\nu_2-1} dx = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \frac{\Gamma(\nu_1 + 1)\Gamma(\nu_2)}{\Gamma(\nu_1 + \nu_2 + 1)} = \frac{\nu_1}{\nu_1 + \nu_2} \end{aligned}$$

Por otra parte,

$$\begin{aligned} \mathbb{E}[X^2] &= \int_{-\infty}^\infty x^2 f(x) dx = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \int_0^1 x^{\nu_1+1} (1-x)^{\nu_2-1} dx \\ &= \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \frac{\Gamma(\nu_1 + 2)\Gamma(\nu_2)}{\Gamma(\nu_1 + \nu_2 + 2)} = \frac{\nu_1(\nu_1 + 1)}{(\nu_1 + \nu_2)(\nu_1 + \nu_2 + 1)} \end{aligned}$$

Finalmente,

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{\nu_1(\nu_1 + 1)}{(\nu_1 + \nu_2)(\nu_1 + \nu_2 + 1)} - \left(\frac{\nu_1}{\nu_1 + \nu_2} \right)^2 \\ &= \frac{\nu_1\nu_2}{(\nu_1 + \nu_2)^2(\nu_1 + \nu_2 + 1)}. \end{aligned}$$

\square

5. Distribución normal.

Sean $\mu \in \mathbb{R}$ y $\sigma > 0$. Se dice que la variable aleatoria X tiene distribución *normal de parámetros* μ, σ^2 , y se denota $X \sim \mathcal{N}(\mu, \sigma^2)$, si X es absolutamente continua con función densidad de probabilidades

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

El cálculo de $\mathbb{E}[X]$ y $\mathbb{V}(X)$ se reduce al caso $X \sim \mathcal{N}(0, 1)$. Para ello, basta observar que $Y \sim \mathcal{N}(\mu, \sigma^2)$ si y solo si $Y = \sigma X + \mu$, donde $X \sim \mathcal{N}(0, 1)$ y usar las identidades $\mathbb{E}[\sigma X + \mu] = \sigma \mathbb{E}[X] + \mu$ y $\mathbb{V}(\sigma X + \mu) = \sigma^2 \mathbb{V}(X)$. En lo que sigue suponemos que $X \sim \mathcal{N}(0, 1)$ y denotamos su densidad mediante

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Es evidente que $\mathbb{E}[X] = 0$. En consecuencia,

$$\mathbb{V}(X) = \mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 \varphi(x) dx$$

Observando que $\varphi'(x) = -x\varphi(x)$ e integrando por partes se obtiene,

$$\mathbb{V}(X) = \int_{-\infty}^{\infty} x(x\varphi(x)) dx = -x\varphi(x) \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \varphi(x) dx = 0 + 1.$$

□

7. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Bertsekas, D. P., Tsitsiklis, J. N.: Introduction to Probability. M.I.T. Lecture Notes. (2000)
2. Billingsley, P.: Probability and Measure. John Wiley & Sons, New York. (1986)
3. Durrett, R. Elementary Probability for Applications. Cambridge University Press, New York. (2009)
4. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons, New York. (1957)
5. Kolmogorov, A. N.: The Theory of Probability. Mathematics. Its Content, Methods, and Meaning. Vol 2. The M.I.T. Press, Massachusetts. (1963) pp. 229-264.
6. Ross, S.: Introduction to Probability and Statistics for Engineers and Scientists. Academic Press, San Diego. (2004)
7. Ross, S.: Introduction to Probability Models. Academic Press, San Diego. (2007)
8. Soong, T. T.: Fundamentals of Probability and Statistics for Engineers. John Wiley & Sons Ltd. (2004)

Transformaciones de variables aleatorias (Borradores, Curso 23)

Sebastian Grynberg

3 de abril de 2013



*Mi unicornio azul ayer se me perdió,
pastando lo dejé y desapareció.*
(Silvio Rodríguez)

Índice

1. Funciones de variables aleatorias	2
1.1. Método básico: eventos equivalentes	2
1.2. Funciones a trozos: dividir y conquistar	5
1.3. Funciones inyectivas suaves	6
1.4. Funciones suaves	7
2. Funciones de vectores aleatorios	7
2.1. Método básico: eventos equivalentes	7
2.1.1. Suma de variables	9
2.1.2. Mínimo	10
2.2. El método del Jacobiano	10
2.3. Funciones k a 1	15
3. Mínimo y máximo de dos exponenciales independientes	18
4. Funciones regulares e independencia	19
5. Bibliografía consultada	20

1. Funciones de variables aleatorias

Sea X una variable aleatoria definida sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Sea $g : D \subseteq \mathbb{R} \rightarrow \mathbb{R}$ una función cuyo dominio D contiene al rango de X : $X(\Omega) := \{x(\omega) : \omega \in \Omega\}$. Entonces $Y = g(X)$ está bien definida y será una variable aleatoria si y sólo si

$$\{\omega \in \Omega : g(X) \leq y\} \in \mathcal{A} \quad \text{para todo } y \in \mathbb{R}. \quad (1)$$

En palabras, si $g^{-1}((-\infty, y]) := \{x \in \mathbb{R} : g(x) \leq y\}$, el conjunto $\{X \in g^{-1}(-\infty, y]\}$ debe tener asignada probabilidad. Este es típicamente el caso. Por ejemplo, si X es discreta, cualquier función g cuyo dominio contenga al rango de X satisface (1). Si X no es discreta, cualquier función g seccionalmente continua cuyo dominio contenga al rango de X satisface (1).

1.1. Método básico: eventos equivalentes

Si queremos hallar la función de distribución de $Y = g(X)$ tenemos que calcular

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \in g^{-1}(-\infty, y]). \quad (2)$$

Los siguientes ejemplos ilustran el *método básico* para hacerlo.

Ejemplo 1.1 (Del péndulo a la distribución de Cauchy). Sea Θ el ángulo de un péndulo medido desde la vertical cuyo extremo superior se encuentra sostenido del punto $(0, 1)$. Sea $(X, 0)$ el punto de intersección de la recta que contiene al péndulo y el eje x -ver la Figura 1-. Trigonometría mediante, sabemos que

$$X = \tan \Theta$$

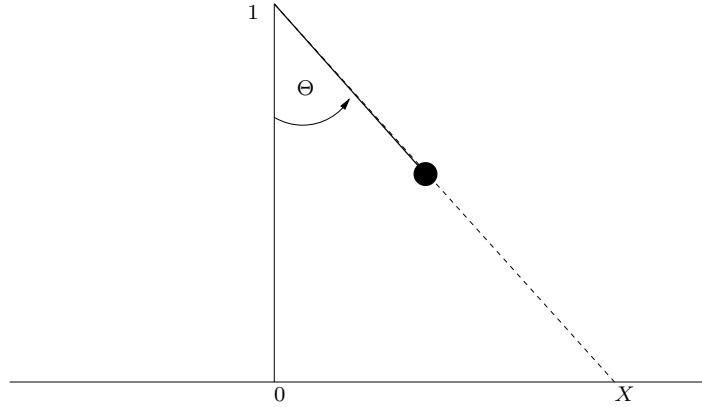


Figura 1: Péndulo.

Si el ángulo Θ es una variable aleatoria uniformemente distribuida sobre el intervalo $(-\frac{\pi}{2}, \frac{\pi}{2})$, cuál es la distribución de X ?

Primero observamos que para cada $\theta \in (-\pi/2, \pi/2)$ tenemos que

$$\mathbb{P}(\Theta \leq \theta) = \frac{\theta - (-\pi/2)}{\pi/2 - (-\pi/2)} = \frac{\theta + \pi/2}{\pi} = \frac{1}{2} + \frac{\theta}{\pi}.$$

De allí se deduce que

$$\mathbb{P}(X \leq x) = \mathbb{P}(\tan \Theta \leq x) = \mathbb{P}(\Theta \leq \arctan x) = \frac{1}{2} + \frac{1}{\pi} \arctan x,$$

y derivando obtenemos que

$$f_X(x) = \frac{1}{\pi(1+x^2)}.$$

□

Teorema 1.2. *Sea X una variable aleatoria continua con función de distribución creciente. Entonces, $Y = F_X(X) \sim \mathcal{U}(0, 1)$.*

Demostración. El análisis se reduce a examinar el comportamiento de la función de distribución de Y sobre el intervalo $(0, 1)$. Para cada $y \in (0, 1)$ vale que

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(F_X(X) \leq y) = \mathbb{P}(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

□

Corolario 1.3. *Sea X una variable aleatoria continua con función de distribución creciente. Sea Y una variable aleatoria cualquiera. Entonces X puede transformarse en una copia de Y haciendo lo siguiente: $\hat{Y} = F_Y^{-1}(F_X(X))$, donde F_Y^{-1} es la inversa generalizada de Y .*

Ejemplo 1.4. Construir una moneda equilibrada X usando una variable aleatoria T con distribución exponencial de intensidad 1.

$$\hat{X} = \mathbf{1} \left\{ \frac{1}{2} < 1 - e^{-T} < 1 \right\}.$$

□

El siguiente ejemplo puede considerarse un prototipo que ilustra cómo tratar con las funciones de variables aleatorias cuando no son inyectivas.

Ejemplo 1.5 (Prototipo). Sea X una variable aleatoria cualquiera y sea $Y = X^2$. Queremos determinar la distribución de Y .

1. Cálculo explícito de la función de distribución. La función de distribución de Y se calcula observando que $g(x) = x^2$ y utilizando la fórmula: $F_Y(y) = \mathbb{P}(X \in g^{-1}((-\infty, y]))$. En este caso, el conjunto $g^{-1}((-\infty, y])$ adopta la forma

$$g^{-1}((-\infty, y]) = \{x \in \mathbb{R} : x^2 \leq y\} = \begin{cases} [-\sqrt{y}, \sqrt{y}] & \text{si } y \geq 0, \\ \emptyset & \text{si } y < 0. \end{cases}$$

Por lo tanto,

$$F_Y(y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \mathbf{1}\{y \geq 0\} = (F_X(\sqrt{y}) - F_X(-\sqrt{y})) \mathbf{1}\{y \geq 0\}. \quad (3)$$

En particular, si X es continua, $\mathbb{P}(X = x) = 0$ para todo $x \in \mathbb{R}$ y la identidad (3) adopta la forma

$$F_Y(y) = (F_X(\sqrt{y}) - F_X(-\sqrt{y})) \mathbf{1}\{y > 0\}. \quad (4)$$

2. Cálculo explícito de la densidad de probabilidades. Si X es absolutamente continua con densidad de probabilidades $f_X(x)$, la densidad de probabilidades de $Y = X^2$ se obtiene derivando la función de distribución $F_Y(y)$. De la identidad (4) se deduce que:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \left(f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} - f_X(-\sqrt{y}) \frac{1}{-2\sqrt{y}} \right) \mathbf{1}\{y > 0\} \\ &= \frac{1}{2\sqrt{y}} (f_X(\sqrt{y}) + f_X(-\sqrt{y})) \mathbf{1}\{y > 0\}. \end{aligned} \quad (5)$$

□

Ejemplo 1.6 (De continua a discreta). Sea $U \sim \mathcal{U}(0, 1)$. Hacemos $Y = [10U]$, donde $[x]$ representa la parte entera de $x \in \mathbb{R}$. Queremos determinar la función de probabilidad de Y .

En primer lugar observamos que la variable aleatoria Y es el primer dígito decimal de un número elegido al azar sobre el intervalo $(0, 1)$. Los posibles valores de Y son $0, 1, \dots, 9$. Para cada $y \in \{0, 1, \dots, 9\}$ vale que

$$\mathbb{P}(Y = y) = \mathbb{P}\left(\frac{y}{10} < U \leq \frac{y+1}{10}\right) = \frac{1}{10}.$$

En otras palabras, $Y \sim \mathcal{U}\{0, 1, \dots, 9\}$.

□

Ejemplo 1.7. Sea $T \sim \text{Exp}(\lambda)$ la duración en minutos de una llamada telefónica. Se factura un pulso cada t_0 minutos o fracción. Queremos determinar la distribución de la cantidad de pulsos facturados por la llamada.

La cantidad de pulsos facturados por la llamada se describe por:

$$N = \sum_{n \geq 1} n \mathbf{1}\{(n-1)t_0 < T \leq nt_0\}.$$

Notando que $N > n \iff T > nt_0$ obtenemos que

$$P(N > n) = e^{-\lambda nt_0} = \left(e^{-\lambda t_0}\right)^n = \mathbb{P}(T > t_0)^n.$$

Por lo tanto, $N \sim \text{Geométrica}(\mathbb{P}(T \leq t_0))$.

□

Ejemplo 1.8 (Variables discretas). Sea X una variable aleatoria discreta a valores $(x_i)_{i \geq 1}$. De la relación $Y = g(X)$ se deduce que los posibles valores de Y son $y_i = g(x_i)$, $i \geq 1$. Si la función de probabilidad de X está dada por $p_X(x_i) = p_i$, $i \geq 1$, la función de probabilidad de Y se determina por

$$p_Y(y_i) = \mathbb{P}(Y = y_i) = \mathbb{P}(X \in g^{-1}(y_i)) = \sum_{x \in g^{-1}(y_i)} p_x.$$

□

Ejercicios adicionales

1. Sea X una variable aleatoria discreta tal que $P(X = -1) = 1/2$, $\mathbb{P}(X = 0) = 1/4$ y $\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = 1/8$. Hallar la función de probabilidad de Y para $Y = 2X + 1$ y para $Y = 2X^2 + 1$.

1.2. Funciones a trozos: dividir y conquistar

Sea X una variable y sea A_1, A_2, \dots una partición de \mathbb{R} tal que $\mathbb{P}(X \in A_i) > 0$ para todo $i \geq 1$. Consideramos una función a trozos definida por

$$g(x) = \sum_{i \geq 1} g_i(x) \mathbf{1}\{x \in A_i\},$$

donde, para cada $i \geq 1$, $g_i : \mathbb{R} \rightarrow \mathbb{R}$, es una función tal que $g_i(X)$ es una variable aleatoria. Si se quiere hallar la distribución de

$$Y = g(X) = \sum_{i \geq 1} g_i(X) \mathbf{1}\{X \in A_i\}$$

se puede hacer lo siguiente: considerar las variables truncadas $X_i = X | X \in A_i$, hallar las distribuciones de las variables $Y_i = g_i(X_i)$ y luego ponderarlas con los pesos $\mathbb{P}(X \in A_i)$:

$$F_Y(y) = \sum_{i \geq 1} F_{Y_i}(y) \mathbb{P}(X \in A_i). \quad (6)$$

En efecto, por una parte tenemos que

$$\begin{aligned} F_Y(y) &= \mathbb{P} \left(\sum_{j \geq 1} g_j(X) \mathbf{1}\{X \in A_j\} \leq y \right) = \sum_{i \geq 1} \mathbb{P} \left(\sum_{j \geq 1} g_j(X) \mathbf{1}\{X \in A_j\} \leq y, X \in A_i \right) \\ &= \sum_{i \geq 1} \mathbb{P}(g_i(X) \leq y, X \in A_i) = \sum_{i \geq 1} \mathbb{P}(X \in g_i^{-1}(-\infty, y] \cap A_i). \end{aligned} \quad (7)$$

Por otra parte,

$$F_{Y_i}(y) = \mathbb{P}(g_i(X_i) \leq y) = \mathbb{P}(X_i \in g_i^{-1}(-\infty, y]) = \frac{\mathbb{P}(X \in g_i^{-1}(-\infty, y] \cap A_i)}{\mathbb{P}(X \in A_i)}.$$

Equivalentemente,

$$P(X \in g_i^{-1}(-\infty, y] \cap A_i) = F_{Y_i}(y) \mathbb{P}(X \in A_i). \quad (8)$$

Combinando (7) y (8) se obtiene (6).

□

1.3. Funciones inyectivas suaves

Teorema 1.9 (Cambio de variables). Sea X una variable aleatoria absolutamente continua con densidad de probabilidades $f_X(x)$. Sea $Y = g(X)$, donde g es una función monótona con derivada no nula. Entonces Y es absolutamente continua y admite una densidad de probabilidades de la forma

$$f_Y(y) = \frac{f_X(x)}{|g'(x)|} \Big|_{x=g^{-1}(y)}. \quad (9)$$

Demostración.

1. La función g es creciente: $g(x_1) \leq g(x_2)$ para $x_1 \leq x_2$. En tal caso la función inversa g^{-1} también es creciente. En consecuencia,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)). \quad (10)$$

La función $F_Y(y)$ es derivable porque es una composición de funciones derivables. Derivando con respecto a y y usando la regla de la cadena se obtiene

$$\frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}.$$

2. La función g es decreciente: $g(x_1) \geq g(x_2)$ para $x_1 \leq x_2$. En este caso la función inversa g^{-1} también es decreciente. En consecuencia,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)). \quad (11)$$

Derivando con respecto a y se obtiene

$$\frac{d}{dy} F_Y(y) = \frac{d}{dy} (1 - F_X(g^{-1}(y))) = -\frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}.$$

□

Corolario 1.10 (Cambio lineal). Dados $a > 0$ y $b \in \mathbb{R}$, la densidad de probabilidades de $Y = aX + b$ adopta la forma

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right). \quad (12)$$

En palabras, desde el punto de vista de la densidad de probabilidades, el cambio lineal $y = ax + b$ efectúa una *traslación en b* seguida de un *cambio de escala de 1 en a* sobre la densidad original. Cuando el parámetro a se achica, los valores de Y tienden a estar más concentrados (alrededor del valor medio) y cuando a se agranda, tienden a dispersarse. □

Ejemplo 1.11 (Variables exponenciales). Se dice que la variable aleatoria Y tiene *distribución exponencial de intensidad* $\lambda > 0$, y se denota $Y \sim \text{Exp}(\lambda)$, si $Y = \frac{1}{\lambda}X$, donde X es una variable aleatoria absolutamente continua que admite una densidad de probabilidades de la forma $f_X(x) = e^{-x}\mathbf{1}\{x \geq 0\}$. De (12) se deduce que Y admite una densidad de probabilidades de la forma $f_Y(y) = \lambda e^{-\lambda y}\mathbf{1}\{y \geq 0\}$. □

Ejemplo 1.12 (Variables Normales). Sean $\mu \in \mathbb{R}$ y $\sigma > 0$. Se dice que la variable aleatoria Y tiene distribución *normal de parámetros* μ, σ^2 , y se denota $Y \sim \mathcal{N}(\mu, \sigma^2)$, si $Y = \sigma X + \mu$, donde X es una variable aleatoria absolutamente continua con densidad de probabilidades $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. De (12) se deduce que Y admite una densidad de probabilidades de la forma $f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$. □

1.4. Funciones suaves

Nota Bene. Las fórmulas (10) y (11) permiten calcular explícitamente la función de distribución, F_Y , para transformaciones monótonas (continuas) $Y = g(X)$, independientemente de la clase de variable que sea X . ¿Qué hacer cuando la transformación g es suave pero no es inyectiva?

Ejemplo 1.13. Sea $X \sim \mathcal{N}(0, 1)$. Segundo la fórmula (5) la densidad de probabilidades de $Y = X^2$ es $f_Y(y) = \frac{1}{2\sqrt{y}} (\varphi(\sqrt{y}) + \varphi(-\sqrt{y})) \mathbf{1}\{y > 0\}$, donde $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Por lo tanto,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} \mathbf{1}\{y > 0\}.$$

En otras palabras, si $X \sim \mathcal{N}(0, 1)$, entonces $X^2 \sim \Gamma(1/2, 1/2)$. □

El Teorema 1.9 puede generalizarse del siguiente modo

Teorema 1.14 (Cambio de variables II). Sea X una variable aleatoria absolutamente continua con densidad de probabilidades $f_X(x)$. Sea $Y = g(X)$, donde g es una función derivable con derivada no nula (salvo en contables puntos). Si para cada $y \in \mathbb{R}$, el conjunto $g^{-1}(y) = \{x \in \mathbb{R} : g(x) = y\}$ es discreto, entonces Y es absolutamente continua y admite una función densidad de probabilidades de la forma

$$f_Y(y) = \sum_{x \in g^{-1}(y)} \frac{f_X(x)}{|g'(x)|}.$$

Se sobreentiende que si $g^{-1}(y) = \emptyset$, $f_Y(y) = 0$.

Ejercicios adicionales

2. [James p.98] Si X tiene densidad $f_X(x)$, cuál es la densidad de $Y = \cos X$?

2. Funciones de vectores aleatorios

2.1. Método básico: eventos equivalentes

Sea $\mathbf{X} = (X_1, \dots, X_n)$ un vector aleatorio definido sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ y sea $g : \mathbb{R}^n \rightarrow \mathbb{R}$ una función cualquiera. Entonces, $Y := g(\mathbf{X})$ será una variable aleatoria si y solo si $\{\omega \in \Omega : g(\mathbf{X}(\omega)) \leq y\} \in \mathcal{A}$ para todo $y \in \mathbb{R}$. La función de distribución de Y , $F_Y(y)$, se puede calcular mediante la función de distribución de \mathbf{X} de la siguiente manera:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(\mathbf{X}) \leq y) = \mathbb{P}(\mathbf{X} \in \mathcal{B}_y), \quad (13)$$

donde $\mathcal{B}_y := g^{-1}((-\infty, y]) = \{\mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}) \leq y\}$.

Caso bidimensional continuo. Sea (X, Y) un vector aleatorio con densidad conjunta $f_{X,Y}(x, y)$. Cualquier función continua a valores reales $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ define una nueva variable aleatoria $Z := g(X, Y)$. La función de distribución de Z , $F_Z(z) = \mathbb{P}(Z \leq z)$, se puede obtener a partir de la densidad conjunta de X e Y de la siguiente forma:

1. Para cada $z \in \mathbb{R}$ se determina el conjunto $\mathcal{B}_z \subset \mathbb{R}^2$ de todos los puntos (x, y) tales que $g(x, y) \leq z$.
2. Integrando la densidad conjunta $f_{X,Y}(x, y)$ sobre el conjunto \mathcal{B}_z se obtiene la función de distribución de Z :

$$F_Z(z) = \iint_{\mathcal{B}_z} f_{X,Y}(x, y) dx dy. \quad (14)$$

3. La densidad de Z se obtiene derivando la función de distribución respecto de z . □

Ejemplo 2.1. Sean X e Y dos variables aleatorias independientes cada una con distribución uniforme sobre el intervalo $[-1, 1]$. Se quiere hallar la función de distribución y la densidad de $Z = |X - Y|$.

La función de distribución de la variable $Z = |X - Y|$ se puede obtener observando la Figura 2.

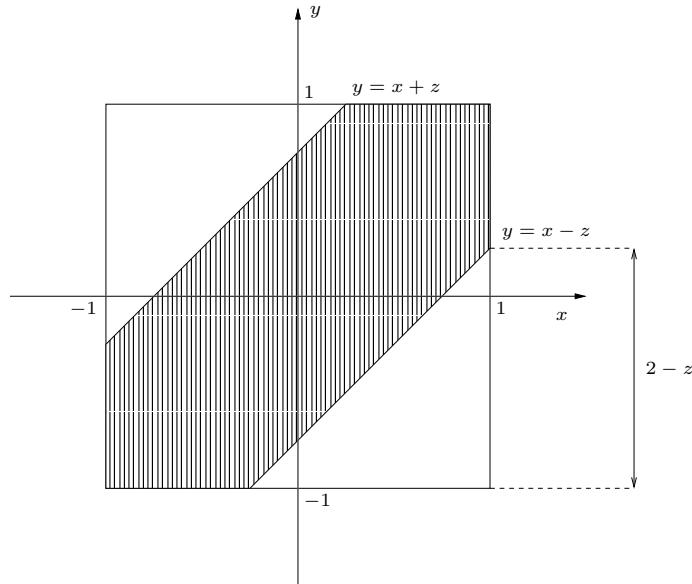


Figura 2: La región sombreada representa los puntos del cuadrado $[-1, 1] \times [-1, 1]$ tales que $|x - y| \leq z$, $0 \leq z \leq 2$ y su área es $4 - (2 - z)^2 = 4z - z^2$.

Debido a que las variables aleatorias X e Y son independientes y uniformemente distribuidas sobre el intervalo $[-1, 1]$, tenemos que $\mathbb{P}((X, Y) \in \mathcal{B}) = \text{área}(\mathcal{B})/4$, para cualquier región \mathcal{B} contenida en el cuadrado $[-1, 1] \times [-1, 1]$ para la que tenga sentido la noción de área. En consecuencia, $F_Z(z) = P(|X - Y| \leq z) = (4z - z^2)/4$ para todo $z \in [0, 2]$. Derivando esta última expresión respecto de z se obtiene la densidad de $Z = |X - Y|$: $f_Z(z) = (\frac{2-z}{2}) \mathbf{1}\{z \in (0, 2)\}$. □

Caso bidimensional discreto. Sea (X, Y) un vector aleatorio discreto sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, con función de probabilidad conjunta $p_{X,Y}(x, y)$. Sea $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ una función cualquiera, $Z := g(X, Y)$ es una nueva variable aleatoria, cuya función de probabilidad, $p_Z(z)$, se obtiene de la siguiente manera:

$$p_Z(z) = \mathbb{P}(Z = z) = \mathbb{P}(g(X, Y) = z) = \sum_{(x,y) \in \mathcal{B}_z} p_{X,Y}(x, y), \quad (15)$$

donde $\mathcal{B}_z = \{(x, y) \in X(\Omega) \times Y(\Omega) : g(x, y) = z\}$. \square

2.1.1. Suma de variables

Ejemplo 2.2 (Suma). Sean X, Y dos variables aleatorias con densidad conjunta $f_{X,Y}(x, y)$ y sea $Z = X + Y$. Para cada $z \in \mathbb{R}$, $\mathcal{B}_z = \{(x, y) \in \mathbb{R}^2 : y \leq z - x\}$. Usando la fórmula (14) se obtiene la función de distribución de Z

$$F_Z(z) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-x} f_{X,Y}(x, y) dy \right) dx. \quad (16)$$

La densidad de Z se obtiene derivando respecto de z la función de distribución $F_Z(z)$

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x) dx. \quad (17)$$

Ejemplo 2.3 (Suma de variables independientes). Sean X, Y dos variables aleatorias continuas e independientes con densidad conjunta $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. Según la fórmula (17) la densidad de probabilidades de la suma $Z = X + Y$ es

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x) dx = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x) dx \quad (18)$$

y se denomina el *producto convolución*, $f_X * f_Y$, de las densidades marginales f_X y f_Y .

Si las densidades marginales $f_X(x)$ y $f_Y(y)$ concentran la masa en $[0, \infty)$ la fórmula (18) del producto convolución es un poco más sencilla:

$$(f_X * f_Y)(z) = \int_0^{\infty} f_X(x)f_Y(z-x) dx = \int_0^z f_X(x)f_Y(z-x) dx. \quad (19)$$

\square

Ejemplo 2.4 (Suma de exponenciales independientes de igual intensidad). Sean X e Y variables aleatorias independientes con distribución exponencial de intensidad $\lambda > 0$. La densidad de la suma $X + Y$ es

$$f_{X+Y}(z) = \int_0^z \lambda e^{-\lambda x} \lambda e^{z-x} dx = \lambda^2 z e^{-\lambda z}. \quad (20)$$

En el lado derecho de la identidad (20) se puede reconocer la densidad de la distribución Gamma: $\Gamma(2, \lambda)$. \square

2.1.2. Mínimo

Queremos caracterizar la función de distribución del mínimo entre dos variables aleatorias X e Y , $U := \min\{X, Y\}$. En primer lugar observamos que para cada $u \in \mathbb{R}$ vale que

$$\begin{aligned} F_U(u) &= \mathbb{P}(U \leq u) = \mathbb{P}(\min\{X, Y\} \leq u) = 1 - \mathbb{P}(\min\{X, Y\} > u) \\ &= 1 - \mathbb{P}(X > u, Y > u). \end{aligned} \quad (21)$$

Si (X, Y) es continuo con función de densidad conjunta $f_{X,Y}(x, y)$ tenemos que

$$F_U(u) = 1 - \int_u^\infty \int_u^\infty f_{X,Y}(x, y) dx dy. \quad (22)$$

Si (X, Y) es discreto con función de probabilidad conjunta $p_{X,Y}(x, y)$ tenemos que

$$F_U(u) = 1 - \sum_{x>u} \sum_{y>u} p_{X,Y}(x, y). \quad (23)$$

Si X e Y son independientes tenemos que

$$F_U(u) = 1 - \mathbb{P}(X > u)\mathbb{P}(Y > u). \quad (24)$$

Etcétera...

Ejemplo 2.5 (Mínimo de exponentiales independientes). Sean X_1 e X_2 variables aleatorias exponentiales independientes de intensidades λ_1 y λ_2 respectivamente. De acuerdo con la identidad (24) tenemos que la función de distribución del mínimo $U = \min\{X_1, X_2\}$ es

$$F_U(u) = (1 - e^{-\lambda_1 u})e^{-\lambda_2 u}\mathbf{1}\{u \geq 0\} = (1 - e^{-(\lambda_1 + \lambda_2)u})\mathbf{1}\{u \geq 0\}. \quad (25)$$

En palabras, *el mínimo de dos variables exponentiales independientes es una exponencial cuya intensidad es la suma de las intensidades de las variables originales.*

□

2.2. El método del Jacobiano

Teorema 2.6 (Cambio de variables en la integral múltiple). *Sea $f : \mathbb{R}^n \rightarrow \mathbb{R}$ una función integrable. Sean $G_0 \subset \mathbb{R}^n$ y $G \subset \mathbb{R}^n$ regiones abiertas y sea $h : G_0 \rightarrow G$, $h = (h_1, \dots, h_n)$ una biyección entre G_0 y G , cuyas componentes tienen derivadas parciales de primer orden continuas. Esto es, para todo $1 \leq i, j \leq n$, las funciones $\frac{\partial h_i(\mathbf{y})}{\partial y_j}$ son continuas. Si el Jacobiano de h es diferente de cero en casi todo punto, entonces,*

$$\int_A f(\mathbf{x}) d\mathbf{x} = \int_{h^{-1}(A)} f(h(\mathbf{y})) |J_h(\mathbf{y})| d\mathbf{y},$$

para todo conjunto abierto $A \subset G$, donde

$$J_h(\mathbf{y}) = \det \left(\left(\frac{\partial h_i(\mathbf{y})}{\partial y_j} \right)_{i,j} \right).$$

El siguiente resultado, que caracteriza la distribución de un cambio de variables aleatorias, es una consecuencia inmediata del Teorema 2.6.

Corolario 2.7. Sea \mathbf{X} un vector aleatorio n -dimensional con función densidad de probabilidad $f_{\mathbf{X}}(\mathbf{x})$. Sean $G_0 \subset \mathbb{R}^n$ y $G \subset \mathbb{R}^n$ regiones abiertas y sea $g : G \rightarrow G_0$ una biyección cuya función inversa $h = g^{-1}$ satisface las hipótesis del Teorema 2.6. Si $\mathbb{P}(\mathbf{X} \in G) = 1$, entonces, el vector aleatorio $\mathbf{Y} = g(\mathbf{X})$ tiene función densidad de probabilidad $f_{\mathbf{Y}}(\mathbf{y})$ de la forma:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y}))|J_{g^{-1}}(\mathbf{y})|. \quad (26)$$

Demostración. Cualquiera sea el conjunto abierto $B \subset G_0$ tenemos

$$\mathbb{P}(\mathbf{Y} \in B) = \mathbb{P}(g(\mathbf{X}) \in B) = \mathbb{P}(\mathbf{X} \in g^{-1}(B)) = \int_{g^{-1}(B)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Poniendo $f = f_{\mathbf{X}}$ y $h = g^{-1}$ en el Teorema 2.6 se obtiene

$$\int_{g^{-1}(B)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int_B f_{\mathbf{X}}(g^{-1}(\mathbf{y}))|J_{g^{-1}}(\mathbf{y})| d\mathbf{y}.$$

En consecuencia,

$$\mathbb{P}(\mathbf{Y} \in B) = \int_B f_{\mathbf{X}}(g^{-1}(\mathbf{y}))|J_{g^{-1}}(\mathbf{y})| d\mathbf{y}.$$

Por lo tanto, el vector aleatorio \mathbf{Y} tiene función densidad de probabilidad de la forma $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y}))|J_{g^{-1}}(\mathbf{y})|$. \square

Nota Bene. Operativamente, la fórmula (26) para hallar la densidad conjunta de $\mathbf{Y} = g(\mathbf{X})$ involucra los siguientes pasos: 1. Invertir las variables (i.e., despejar las x 's en función de las y 's). 2. Calcular el Jacobiano de la inversa de g (i.e., calcular el determinante de la matriz formada por las derivadas parciales de las x_i respecto de las y_j). 3. Substituir los resultados obtenidos en los pasos 1. y 2. en la fórmula (26). **Aunque mecánico, el método del jacobiano es un método de naturaleza analítica muy poderoso.** \square

Nota Bene. Con frecuencia es más fácil obtener el jacobiano de \mathbf{y} en relación a \mathbf{x} , pues \mathbf{Y} es una función de \mathbf{X} . Hay que recordar que los dos jacobianos son recíprocos y que $J_{g^{-1}}(\mathbf{y})$ se puede obtener a partir de $J_g(\mathbf{x})$, invirtiendo este último y substituyendo \mathbf{x} por $g^{-1}(\mathbf{y})$. Esta regla es análoga a la regla para la derivada de una función inversa en el caso unidimensional:

$$\frac{dg^{-1}(y)}{dy} = \frac{1}{g'(x)} \Big|_{x=g^{-1}(y)} = \frac{1}{g'(g^{-1}(y))}.$$

\square

Ejemplo 2.8 (Transformaciones lineales). Si $(X_1, X_2) = (aY_1 + bY_2, cY_1 + dY_2)$. Entonces,

$$f_{Y_1, Y_2}(y_1, y_2) = |ad - bc| f_{X_1, X_2}(ay_1 + by_2, cy_1 + dy_2).$$

En general, si $\mathbf{X} = A\mathbf{Y}$, donde $A \in \mathbb{R}^{n \times n}$ es una matriz inversible, se obtiene

$$f_{\mathbf{Y}}(\mathbf{y}) = |\det(A)| f_{\mathbf{X}}(A\mathbf{y}). \quad (27)$$

Ejemplo 2.9 (Suma y resta de normales independientes). Sean X_1 y X_2 dos variables aleatorias independientes con distribuciones normales $\mathcal{N}(\mu_1, \sigma^2)$ y $\mathcal{N}(\mu_2, \sigma^2)$, respectivamente. Su densidad conjunta es

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}((x_1 - \mu_1)^2 + (x_2 - \mu_2)^2)\right) \quad (28)$$

Consideramos el cambio de variables $(y_1, y_2) = g(x_1, x_2) = (x_1 + x_2, x_1 - x_2)$ cuya inversa es $(x_1, x_2) = g^{-1}(y_1, y_2) = \frac{1}{2}(y_1 + y_2, y_1 - y_2)$. De acuerdo con la fórmula (27) tenemos que

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= \frac{1}{4\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\left(\left(\frac{y_1 + y_2}{2} - \mu_1\right)^2 + \left(\frac{y_1 - y_2}{2} - \mu_2\right)^2\right)\right) \\ &\propto \exp\left(-\frac{1}{4\sigma^2}(y_1^2 - 2(\mu_1 + \mu_2)y_1)\right) \exp\left(-\frac{1}{4\sigma^2}(y_2^2 - 2(\mu_1 - \mu_2)y_2)\right) \\ &\propto \exp\left(-\frac{(y_1 - (\mu_1 + \mu_2))^2}{2(2\sigma^2)}\right) \exp\left(-\frac{(y_2 - (\mu_1 - \mu_2))^2}{2(2\sigma^2)}\right). \end{aligned} \quad (29)$$

De la identidad (29) podemos concluir que las variables Y_1 e Y_2 son independientes y que se distribuyen de la siguiente manera: $Y_1 \sim \mathcal{N}(\mu_1 + \mu_2, 2\sigma^2)$, $Y_2 \sim \mathcal{N}(\mu_1 - \mu_2, 2\sigma^2)$. En otras palabras, *si X_1 y X_2 son dos variables aleatorias independientes con distribuciones normales $\mathcal{N}(\mu_1, \sigma^2)$ y $\mathcal{N}(\mu_2, \sigma^2)$, entonces $X_1 + X_2$ y $X_1 - X_2$ son independientes y $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, 2\sigma^2)$ y $X_1 - X_2 \sim \mathcal{N}(\mu_1 - \mu_2, 2\sigma^2)$* □

Nota Bene. Sean X_1 y X_2 dos variables aleatorias independientes con distribuciones normales $\mathcal{N}(\mu_1, \sigma_1^2)$ y $\mathcal{N}(\mu_2, \sigma_2^2)$, respectivamente. Cálculos similares permiten deducir que $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ y $X_1 - X_2 \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$. Más aún, $X_1 + X_2$ y $X_1 - X_2$ son independientes si y solo si $\sigma_1^2 = \sigma_2^2$. □

Ejemplo 2.10 (Persistencia de la mala suerte). Sean X_1 y X_2 variables aleatorias independientes con distribución común exponencial de intensidad λ . Vamos a hallar la densidad conjunta de (Y_1, Y_2) donde

$$(Y_1, Y_2) = (X_1 + X_2, X_1/X_2).$$

Para ello consideramos la transformación

$$g(x_1, x_2) = (x_1 + x_2, x_1/x_2) = (y_1, y_2).$$

La transformación inversa de g es

$$x_1 = \frac{y_1 y_2}{1 + y_2}, \quad x_2 = \frac{y_1}{1 + y_2} \quad (30)$$

y se obtiene resolviendo un sistema de dos ecuaciones en las variables x_1 y x_2 :

$$\begin{cases} x_1 + x_2 = y_1 \\ x_1/x_2 = y_2 \end{cases} \iff \begin{cases} x_1 + x_2 = y_1 \\ x_1 = y_2 x_2 \end{cases} \iff \begin{cases} (1 + y_2)x_2 = y_1 \\ x_1 = y_2 x_2 \end{cases} \iff \begin{cases} x_2 = \frac{y_1}{1+y_2} \\ x_1 = \frac{y_1 y_2}{1+y_2} \end{cases}$$

El Jacobiano de la transformación inversa $J_{g^{-1}}(y_1, y_2) = \det\left(\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j}\right)$ es

$$\begin{aligned} J_{g^{-1}}(y_1, y_2) &= \frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2} \frac{\partial x_2}{\partial y_1} = \left(\frac{y_2}{1 + y_2}\right) \left(\frac{-y_1}{(1 + y_2)^2}\right) - \left(\frac{y_1}{(1 + y_2)^2}\right) \left(\frac{1}{1 + y_2}\right) \\ &= \frac{-y_1 y_2}{(1 + y_2)^3} - \frac{y_1}{(1 + y_2)^3} = -\frac{y_1(1 + y_2)}{(1 + y_2)^3} = -\frac{y_1}{(1 + y_2)^2}. \end{aligned} \quad (31)$$

Substituyendo los resultados (30) y (31) en la fórmula (26) se obtiene:

$$f_{Y_1, Y_2}(y_1, y_2) = f_{X_1, X_2} \left(\frac{y_1 y_2}{1 + y_2}, \frac{y_1}{1 + y_2} \right) \frac{|y_1|}{(1 + y_2)^2}. \quad (32)$$

Por hipótesis,

$$f_{X_1, X_2}(x_1, x_2) = \lambda e^{-\lambda x_1} \mathbf{1}\{x_1 > 0\} \lambda e^{-\lambda x_2} \mathbf{1}\{x_2 > 0\} = \lambda^2 e^{-\lambda(x_1+x_2)} \mathbf{1}\{x_1 > 0, x_2 > 0\}. \quad (33)$$

De (32) y (33) se obtiene

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= \lambda^2 e^{-\lambda y_1} \frac{y_1}{(1 + y_2)^2} \mathbf{1}\{y_1 > 0, y_2 > 0\} \\ &= \left(\lambda^2 y_1 e^{-\lambda y_1} \mathbf{1}\{y_1 > 0\} \right) \left(\frac{1}{(1 + y_2)^2} \mathbf{1}\{y_2 > 0\} \right). \end{aligned} \quad (34)$$

De (34) se deduce que las variables Y_1 e Y_2 son independientes.

Nota Bene sobre la persistencia de la mala suerte. De (34) se deduce que la densidad del cociente $Y_2 = X_1/X_2$ de dos variables exponenciales independientes de igual intensidad es de la forma

$$f_{Y_2}(y_2) = \frac{1}{(1 + y_2)^2} \mathbf{1}\{y_2 > 0\}. \quad (35)$$

En consecuencia, *la variable Y_2 tiene esperanza infinita*. Se trata de un hecho notable que ofrece una explicación probabilística de un fenómeno conocido por cualquiera que haya entrado en una fila de espera denominado la *persistencia de la mala suerte*¹

¿Por qué? Supongamos que la variable X_1 representa el tiempo de espera para ser atendidos en la fila elegida (a la que llamaremos la fila 1) y que X_2 representa el tiempo de espera en otra fila que estamos observando mientras esperamos ser atendidos (a la que llamaremos la fila 2). El cociente X_1/X_2 representa la proporción del tiempo esperado en la fila 1 en relación al tiempo de espera en fila 2. Por ejemplo, $X_1/X_2 \geq 3$ significa esperamos por lo menos el triple del tiempo que hubiésemos esperado en la otra fila.

Integrando (35) se deduce que

$$\mathbb{P}(Y_2 \leq y_2) = \int_0^{y_2} \frac{1}{(1 + y)^2} dy = 1 - \frac{1}{1 + y_2} = \frac{y_2}{1 + y_2}, \quad y_2 \geq 0$$

Equivalentemente,

$$\mathbb{P}(Y_2 > y_2) = \frac{1}{1 + y_2}, \quad y_2 \geq 0$$

En particular, la probabilidad de que tengamos que esperar por lo menos el triple del tiempo que hubiésemos esperado en la otra fila es 1/4. Aunque de acuerdo con este modelo, en promedio, la mitad de las veces esperamos menos tiempo que en la otra fila, en la práctica, el fenómeno de la *mala suerte* se ve sobredimensionado porque no le prestamos atención a los tiempos cortos de espera.

¹Basta elegir una fila en las múltiples cajas de un supermercado para sufrir este fenómeno y observar que en la fila elegida el tiempo de espera es el doble o el triple que el tiempo de espera en las otras filas.

Para percibir qué significa el resultado $\mathbb{E}[X_1/X_2] = +\infty$ basta simular algunos valores de la variable X_1/X_2 . Por ejemplo, en 10 simulaciones obtuvimos la siguiente muestra:

1.2562, 0.8942, 0.9534, 0.3596, **29.3658**, 1.2641, 3.3443, 0.3452, **13.5228**, 7.1701.

El lector puede extraer sus propias conclusiones. \square

Ejemplo 2.11 (Gammas y Betas). Sean X_1 y X_2 variables aleatorias independientes con distribuciones $\Gamma(\nu_1, \lambda)$ y $\Gamma(\nu_2, \lambda)$. Vamos a hallar la densidad conjunta de (Y_1, Y_2) donde

$$Y_1 = X_1 + X_2, \quad \text{e} \quad Y_2 = \frac{X_1}{X_1 + X_2}.$$

Para ello consideramos la transformación

$$g(x_1, x_2) = \left(x_1 + x_2, \frac{x_1}{x_1 + x_2} \right) = (y_1, y_2).$$

La transformación inversa de g es

$$x_1 = y_1 y_2, \quad x_2 = y_1(1 - y_2). \quad (36)$$

El Jacobiano de la transformación inversa es

$$J_{g^{-1}}(y_1, y_2) = \frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2} \frac{\partial x_2}{\partial y_1} = y_2(-y_1) - y_1(1 - y_2) = -y_1 \quad (37)$$

Substituyendo los resultados (36) y (37) en la fórmula (26) se obtiene:

$$f_{Y_1, Y_2}(y_1, y_2) = f_{X_1, X_2}(y_1 y_2, y_1(1 - y_2)) |y_1|. \quad (38)$$

Por hipótesis,

$$\begin{aligned} f_{X_1, X_2}(x_1, x_2) &= \frac{\lambda^{\nu_1} x_1^{\nu_1-1} e^{-\lambda x_1}}{\Gamma(\nu_1)} \mathbf{1}\{x_1 > 0\} \frac{\lambda^{\nu_2} x_2^{\nu_2-1} e^{-\lambda x_2}}{\Gamma(\nu_2)} \mathbf{1}\{x_2 > 0\} \\ &= \frac{\lambda^{\nu_1+\nu_2} x_1^{\nu_1-1} x_2^{\nu_2-1} e^{-\lambda(x_1+x_2)}}{\Gamma(\nu_1)\Gamma(\nu_2)} \mathbf{1}\{x_1 > 0, x_2 > 0\}. \end{aligned} \quad (39)$$

De (38) y (39) se obtiene

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= \frac{\lambda^{\nu_1+\nu_2} (y_1 y_2)^{\nu_1-1} (y_1(1 - y_2))^{\nu_2-1} e^{-\lambda y_1}}{\Gamma(\nu_1)\Gamma(\nu_2)} \mathbf{1}\{y_1 y_2 > 0, y_1(1 - y_2) > 0\} |y_1| \\ &= \left(\frac{\lambda^{\nu_1+\nu_2} y_1^{\nu_1+\nu_2-1} e^{-\lambda y_1}}{\Gamma(\nu_1 + \nu_2)} \mathbf{1}\{y_1 > 0\} \right) \\ &\quad \times \left(\frac{\Gamma(\nu_1 + \nu_2) y_2^{\nu_1-1} (1 - y_2)^{\nu_2-1}}{\Gamma(\nu_1)\Gamma(\nu_2)} \mathbf{1}\{0 < y_2 < 1\} \right). \end{aligned} \quad (40)$$

Por lo tanto, Y_1 e Y_2 son independientes y sus distribuciones son $Y_1 \sim \Gamma(\nu_1 + \nu_2, \lambda)$, $Y_2 \sim \beta(\nu_1, \nu_2)$:

$$\begin{aligned} f_{Y_1}(y_1) &= \frac{\lambda^{\nu_1+\nu_2}}{\Gamma(\nu_1 + \nu_2)} y_1^{\nu_1+\nu_2-1} e^{-\lambda y_1} \mathbf{1}\{y_1 > 0\}, \\ f_{Y_2}(y_2) &= \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} y_2^{\nu_1-1} (1 - y_2)^{\nu_2-1} \mathbf{1}\{0 < y_2 < 1\}. \end{aligned}$$

\square

Nota Bene. Algunos autores utilizan (y promueven!) el método del Jacobiano como una herramienta para obtener la densidad de variables aleatorias de la forma $Y_1 = g_1(X_1, X_2)$. Hacen lo siguiente: 1. Introducen una variable auxiliar de la forma $Y_2 = g_2(X_1, X_2)$ para obtener un cambio de variables $(g_1, g_2) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. 2. Utilizan la fórmula del Jacobiano (26) para obtener la densidad conjunta de (Y_1, Y_2) a partir de la densidad conjunta de (X_1, X_2) . 3. Obtienen la densidad de Y_1 marginando (i.e., integrando la densidad conjunta de (Y_1, Y_2) con respecto de y_2). Por ejemplo,

Suma: $(X_1, X_2) \rightarrow (X_1 + X_2, X_2) =: (Y_1, Y_2)$. En tal caso, $(x_1, x_2) = (y_1 - y_2, y_2)$ y el Jacobiano tiene la forma $J(y_1, y_2) = \frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2} \frac{\partial x_2}{\partial y_1} = 1$. De donde se obtiene

$$f_{Y_1}(y_1) = \int_{\mathbb{R}} f_{X_1, X_2}(y_1 - y_2, y_2) dy_2.$$

Producto: $(X_1, X_2) \rightarrow (X_1 X_2, X_1) =: (Y_1, Y_2)$. En tal caso, $(x_1, x_2) = (y_2, y_1/y_2)$ y el Jacobiano tiene la forma $J(y_1, y_2) = \frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2} \frac{\partial x_2}{\partial y_1} = -\frac{1}{y_2}$. De donde se obtiene

$$f_{Y_1}(y_1) = \int_{\mathbb{R}} f_{X_1, X_2}(y_2, y_1/y_2) |y_2|^{-1} dy_2.$$

Cociente: $(X_1, X_2) \rightarrow (X_1/X_2, X_2) =: (Y_1, Y_2)$. En tal caso, $(x_1, x_2) = (y_1 y_2, y_2)$ y el Jacobiano tiene la forma $J(y_1, y_2) = \frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2} \frac{\partial x_2}{\partial y_1} = y_2$. De donde se obtiene

$$f_{Y_1}(y_1) = \int_{\mathbb{R}} f_{X_1, X_2}(y_1 y_2, y_2) |y_2| dy_2.$$

□

Ejercicios adicionales

3. [James p.97] Si X, Y, Z tienen densidad conjunta

$$f_{X,Y,Z}(x, y, z) = \frac{6}{(1+x+y+z)^4} \mathbf{1}\{x > 0, y > 0, z > 0\}.$$

Hallar la densidad de la variable aleatoria $W = X + Y + Z$ de dos maneras diferentes (método básico y método del Jacobiano)

2.3. Funciones k a 1

Si la función $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ no es 1 a 1 también podemos utilizar el método del jacobiano para determinar la distribución de $\mathbf{Y} = g(\mathbf{X})$. Basta con que g sea 1 a 1 cuando se la restringe a una de k regiones abiertas disjuntas cuya unión contiene al valor de \mathbf{X} con probabilidad 1.

Supongamos que G, G_1, \dots, G_k son regiones abiertas de \mathbb{R}^n tales que G_1, \dots, G_k son disjuntas dos a dos y que

$$\mathbb{P}\left(\mathbf{X} \in \bigcup_{\ell=1}^k G_\ell\right) = 1.$$

Supongamos además que la restricción de g a G_ℓ , $g|G_\ell$, es una correspondencia 1 a 1 entre G_ℓ y G , para todo $\ell = 1, \dots, k$ y que la función inversa de $g|G_\ell$, denotada por $h^{(\ell)}$, satisface todas las condiciones de la función h del Teorema 2.6.

Teorema 2.12. *Bajo las condiciones enunciadas más arriba, si \mathbf{X} tiene densidad $f_{\mathbf{X}}(\mathbf{x})$, entonces \mathbf{Y} tiene densidad*

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\ell=1}^k f_{\mathbf{X}}(h^{(\ell)}(\mathbf{y})) |J_{h^{(\ell)}}(\mathbf{y})| \mathbf{1}\{\mathbf{y} \in G\}. \quad (41)$$

Demostración. Sea $B \subset G$,

$$\begin{aligned} \mathbb{P}(\mathbf{Y} \in B) &= \mathbb{P}(g(\mathbf{X}) \in B) = \sum_{\ell=1}^k \mathbb{P}(g(\mathbf{X}) \in B, X \in G_{\ell}) = \sum_{\ell=1}^k \mathbb{P}(X \in h^{(\ell)}(B)) \\ &= \sum_{\ell=1}^k \int_{h^{(\ell)}(B)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = (\text{cambio de variables en la integral}) \\ &= \sum_{\ell=1}^k \int_B f_{\mathbf{X}}(h^{(\ell)}(\mathbf{y})) |J_{h^{(\ell)}}(\mathbf{y})| d\mathbf{y} = \int_B \left(\sum_{\ell=1}^k f_{\mathbf{X}}(h^{(\ell)}(\mathbf{y})) |J_{h^{(\ell)}}(\mathbf{y})| \right) d\mathbf{y}. \end{aligned}$$

□

Ejemplo 2.13. Sean X e Y dos variables aleatorias independientes con distribución común $\mathcal{N}(0, 1)$. Mostrar que $Z = X^2 + Y^2$ y $W = X/Y$ son independientes y hallar sus distribuciones.

Solución. La función $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, definida por $g(x, y) = (x^2 + y^2, x/y) = (z, w)$, es 2 a 1.

Sean $G = \{(z, w) : z > 0\}$, $G_1 = \{(x, y) : y > 0\}$, $G_2 = \{(x, y) : y < 0\}$. Entonces, las restricciones $g|G_1$ y $g|G_2$ son correspondencias 1 a 1 entre las regiones abiertas G_i y G , $i = 1, 2$, y $\mathbb{P}((X, Y) \in G_1 \cup G_2) = 1$.

Tenemos que calcular los jacobianos de las funciones inversas $h^{(1)}$ y $h^{(2)}$ en G . Para ello calculamos los jacobianos de las restricciones $g|G_1$ y $g|G_2$, que son los recíprocos de los jacobianos de las inversas, y substituimos el valor (x, y) por el valor $h^{(1)}(z, w)$ o $h^{(2)}(z, w)$. Tenemos

$$J_1(z, w) = \begin{vmatrix} 2x & 2y \\ \frac{1}{y} & -\frac{x}{y^2} \end{vmatrix}^{-1} = \left(-2 \left(\frac{x^2}{y^2} + 1 \right) \right)^{-1} = -\frac{1}{2(w^2 + 1)}$$

y

$$J_2(z, w) = -\frac{1}{2(w^2 + 1)}.$$

Por lo tanto, la densidad de (Z, W) es

$$f_{Z,W}(z, w) = \left(f(h^{(1)}(z, w)) + f(h^{(2)}(z, w)) \right) \frac{1}{2(w^2 + 1)} \mathbf{1}\{(z, w) \in G\}.$$

Como

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2} = \frac{1}{2\pi} e^{-z/2},$$

tenemos

$$f_{Z,W}(z, w) = 2 \left(\frac{1}{2\pi} e^{-z/2} \right) \frac{1}{2(w^2 + 1)} \mathbf{1}\{z > 0, w \in \mathbb{R}\} = \left(\frac{1}{2} e^{-z/2} \mathbf{1}\{z > 0\} \right) \frac{1}{\pi(w^2 + 1)}.$$

Como la densidad conjunta es el producto de dos densidades, concluimos que Z y W son independientes, $Z \sim \text{Exp}(1/2)$ y $W \sim \text{Cauchy}$. □

Ejemplo 2.14 (Mínimo y máximo). Sean X_1, X_2 dos variables aleatorias con densidad conjunta $f_{X_1, X_2}(x_1, x_2)$. Hallar la densidad conjunta de $U = \min(X_1, X_2)$ y $V = \max(X_1, X_2)$.

La función $g(x_1, x_2) = (\min(x_1, x_2), \max(x_1, x_2))$, es 2 a 1.

Sean $G = \{(u, v) : u < v\}$, $G_1 = \{(x_1, x_2) : x_1 < x_2\}$ y $G_2 = \{(x_1, x_2) : x_2 < x_1\}$.

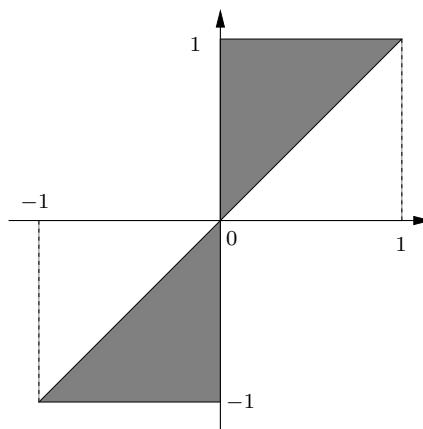
Las restricciones $g|G_1(x_1, x_2) = (x_1, x_2)$ y $g|G_2(x_1, x_2) = (x_2, x_1)$ son correspondencias 1 a 1 entre las regiones abiertas G_i y G , $i = 1, 2$; $\mathbb{P}((X, Y) \in G_1 \cup G_2) = 1$ y los jacobianos de las funciones inversas $h^{(1)}$ y $h^{(2)}$ en G valen 1 y -1, respectivamente. Usando la fórmula (41) obtenemos la densidad conjunta de (U, V) :

$$f_{U,V}(u, v) = (f_{X_1, X_2}(u, v) + f_{X_1, X_2}(v, u)) \mathbf{1}\{u < v\}.$$

□

Ejercicios adicionales

4. La distribución de (X, Y) es uniforme sobre el recinto sombreado



Hallar la densidad conjunta de $(U, V) = (|2Y|, |3X|)$.

5. [James p.99] Sean X_1, \dots, X_n variables aleatorias independientes e idénticamente distribuidas, con densidad común f . Mostrar que la densidad conjunta de

$$U = \min_{1 \leq i \leq n} X_i \quad \text{y} \quad V = \max_{1 \leq i \leq n} X_i$$

es

$$f_{U,V}(u, v) = n(n-1)[F(v) - F(u)]^{n-2} f(u)f(v) \mathbf{1}\{u < v\}.$$

(Sugerencia. Primero hallar $\mathbb{P}(u < U, V \leq v)$. Después, calcular las derivadas parciales cruzadas de la distribución conjunta.)

6. [James p.99] Sean X_1, \dots, X_n variables aleatorias independientes e idénticamente distribuidas, con distribución uniforme sobre el intervalo $[0, 1]$. Sean

$$U = \min_{1 \leq i \leq n} X_i \quad \text{y} \quad V = \max_{1 \leq i \leq n} X_i$$

(a) Mostrar que la densidad conjunta de (U, V) es

$$f_{U,V}(u, v) = n(n-1)(v-u)^{n-2} \mathbf{1}\{0 \leq u < v \leq 1\}.$$

(b) Mostrar que la densidad de $W = V - U$ es

$$f_W(w) = n(n-1)w^{n-2} (1-w) \mathbf{1}\{0 \leq w \leq 1\}.$$

3. Mínimo y máximo de dos exponenciales independientes

Teorema 3.1. Sean X_1 y X_2 dos variables aleatorias independientes con distribuciones exponenciales de intensidades λ_1 y λ_2 respectivamente. Si $U = \min(X_1, X_2)$, $V = \max(X_1, X_2)$, $W = V - U$ y $J = \mathbf{1}\{U = X_1\} + 2\mathbf{1}\{U = X_2\}$, entonces

- (a) $U \sim \text{Exp}(\lambda_1 + \lambda_2)$.
- (b) $\mathbb{P}(J = i) = \lambda_i(\lambda_1 + \lambda_2)^{-1}$, $i = 1, 2$.
- (c) U y J son independientes.
- (d) $f_W(w) = \mathbb{P}(J = 1)f_{X_2}(w) + \mathbb{P}(J = 2)f_{X_1}(w)$.
- (e) U y W son independientes.

Demostración. Primero observamos que para cada $u > 0$ el evento $\{J = 1, U > u\}$ equivale al evento $\{X_2 \geq X_1 > u\}$. En consecuencia,

$$\begin{aligned} \mathbb{P}(J = 1, U > u) &= \int_u^\infty \lambda_1 e^{-\lambda_1 x_1} \left(\int_{x_1}^\infty \lambda_2 e^{-\lambda_2 x_2} dx_2 \right) dx_1 = \int_u^\infty \lambda_1 e^{-\lambda_1 x_1} e^{-\lambda_2 x_1} dx_1 \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2} \int_u^\infty (\lambda_1 + \lambda_2) e^{-(\lambda_1 + \lambda_2)x_1} dx_1 \\ &= \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right) e^{-(\lambda_1 + \lambda_2)u}. \end{aligned} \tag{42}$$

De (42) se deducen (a), (b) y (c).

Si $g : \{(u, v) : 0 < u < v\} \rightarrow \{(u, w) : u > 0, w > 0\}$ es la función definida por $g(u, v) = (u, v - u)$, tenemos que $(U, W) = g(U, V)$. La función g es biyectiva y su inversa $h(u, w) = (u, u + w)$ tiene jacobiano idénticamente igual a 1. Aplicar el método del jacobiano del Corolario 2.7 obtenemos:

$$f_{U,W}(u, w) = f_{U,V}(u, u + w). \tag{43}$$

Por el Ejemplo 2.14 sabemos que la densidad conjunta de U y V es

$$f_{U,V}(u, v) = \lambda_1 \lambda_2 \left(e^{-(\lambda_1 u + \lambda_2 v)} + e^{-(\lambda_1 v + \lambda_2 u)} \right) \mathbf{1}\{0 < u < v\}. \tag{44}$$

Combinando (43) y (44) obtenemos:

$$\begin{aligned}
f_{V,W}(u, w) &= \lambda_1 \lambda_2 \left(e^{-(\lambda_1 u + \lambda_2(u+w))} + e^{-(\lambda_1(u+w) + \lambda_2 w)} \right) \mathbf{1}\{u > 0, w > 0\} \\
&= \lambda_1 \lambda_2 e^{-(\lambda_1 + \lambda_2)u} \left(e^{-\lambda_2 w} + e^{-\lambda_1 w} \right) \mathbf{1}\{u > 0, w > 0\} \\
&= (\lambda_1 + \lambda_2) e^{-(\lambda_1 + \lambda_2)u} \mathbf{1}\{u > 0\} \\
&\quad \times \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \lambda_2 e^{-\lambda_2 w} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \lambda_1 e^{-\lambda_1 w} \right) \mathbf{1}\{w > 0\}.
\end{aligned} \tag{45}$$

De (45) se deducen (d) y (e). \square

Ejercicios adicionales

7. Un avión tiene dos motores cada uno de los cuales funciona durante un tiempo exponencial de media 10 horas independientemente del otro. El avión se mantiene volando mientras funcione alguno de sus motores. Calcular la probabilidad de que el avión se mantenga volando durante más de cinco horas después de que dejó de funcionar un motor.

8. Una cueva será iluminada por dos lámparas L_1 y L_2 cuyas duraciones (en horas) son independientes y tienen distribuciones exponenciales de medias 8 y 10, respectivamente. Sabiendo que desde que se apagó una lámpara la cueva se mantuvo iluminada durante más de una hora calcular la probabilidad de que se haya apagado primero la lámpara L_2 .

4. Funciones regulares e independencia

Definición 4.1. Una función g se dice regular si existen números $\cdots < a_{-1} < a_0 < a_1 < \cdots$, con $a_i \rightarrow \infty$ y $a_{-i} \rightarrow -\infty$, tales que g es continua y monótona sobre cada intervalo (a_i, a_{i+1}) .

Ejemplo 4.2. La función $\sin x$ es regular; todos los polinomios son funciones regulares. Un ejemplo de una función que no es regular es $\mathbf{1}\{x \in \mathbb{Q}\}$. \square

Teorema 4.3. Sean X_1, \dots, X_n variables aleatorias independientes. Si g_1, \dots, g_n son funciones regulares, entonces $g_1(X_1), \dots, g_n(X_n)$ son variables aleatorias independientes.

Demostración. Para simplificar la prueba supondremos que $n = 2$. De la regularidad de las funciones g_1 y g_2 se deduce que para todo $y \in \mathbb{R}$ podemos escribir

$$A_1(y) := \{x : g_1(x) \leq y\} = \cup_i A_{1,i}(y) \quad \text{y} \quad A_2(y) := \{x : g_2(x) \leq y\} = \cup_i A_{2,i}(y),$$

como uniones de intervalos disjuntos dos a dos. Por lo tanto,

$$\begin{aligned}
\mathbb{P}(g_1(X_1) \leq y_1, g_2(X_2) \leq y_2) &= \sum_i \sum_j \mathbb{P}(X_1 \in A_{1,i}(y_1), X_2 \in A_{2,i}(y_2)) \\
&= \sum_i \sum_j \mathbb{P}(X_1 \in A_{1,i}(y_1)) \mathbb{P}(X_2 \in A_{2,i}(y_2)) \\
&= \sum_i \mathbb{P}(X_1 \in A_{1,i}(y_1)) \sum_j \mathbb{P}(X_2 \in A_{2,i}(y_2)) \\
&= \mathbb{P}(g_1(X_1) \leq y_1) \mathbb{P}(g_2(X_2) \leq y_2).
\end{aligned}$$

□

En rigor de verdad, vale un resultado mucho más general.

Teorema 4.4. Si para $1 \leq i \leq n$, $1 \leq j \leq m_i$, $X_{i,j}$ son independientes y $f_i : \mathbb{R}^{m_i} \rightarrow \mathbb{R}$ son medibles entonces $f_i(X_{i,1}, \dots, X_{i,m_i})$ son independientes.

Demostración. Durrett(1996), p.25-27. □

Un caso concreto que usaremos permanentemente al estudiar sumas es el siguiente: si X_1, \dots, X_n son independientes, entonces $X = X_1 + \dots + X_{n-1}$ y X_n son independientes.

Ejercicios adicionales

9. (Fragmentaciones aleatorias.) Si U_1, \dots, U_n son independientes con distribución común $\mathcal{U}(0, 1)$, entonces

$$-\log \prod_{i=1}^n U_i \sim \Gamma(n, 1).$$

10. Una varilla de 1 metro de longitud es sometida a un proceso de fragmentación aleatoria. En la primera fase se elige un punto al azar de la misma y se la divide por el punto elegido en dos varillas de longitudes L_1 y L_2 . En la segunda fase se elige un punto al azar de la varilla de longitud L_1 y se la divide por el punto elegido en dos varillas de longitudes $L_{1,1}$ y $L_{1,2}$. Calcular la probabilidad de que $L_{1,1}$ sea mayor que 25 centímetros.

5. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

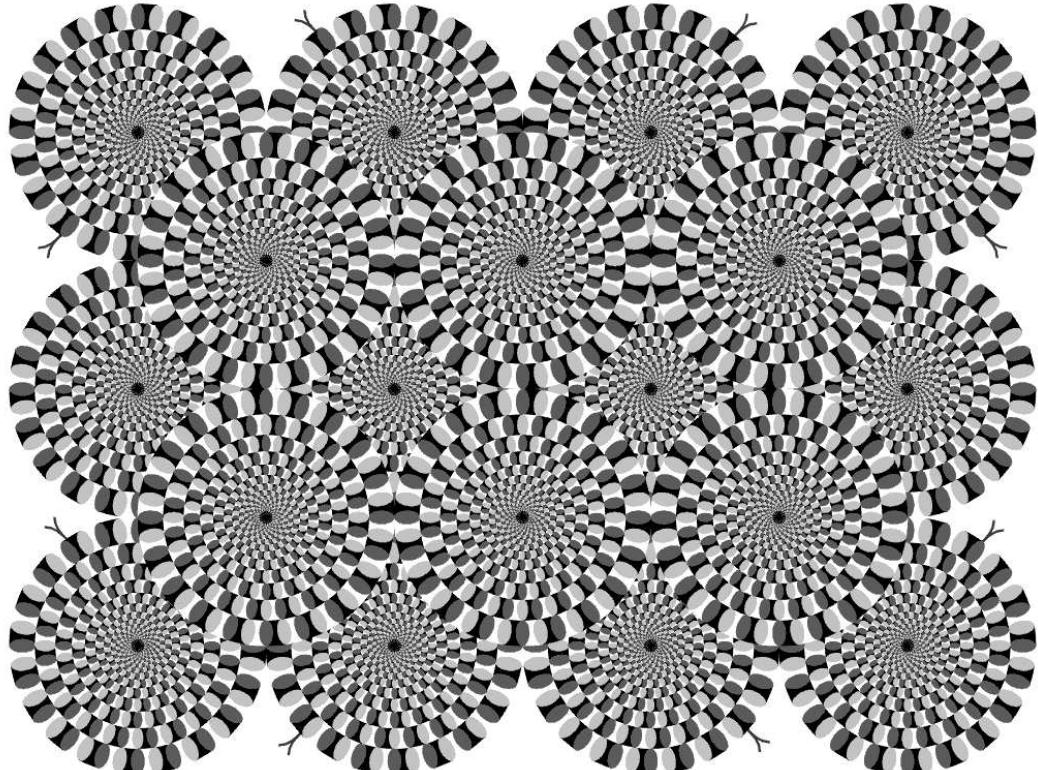
1. Durrett R.: Probability. Theory and Examples. Duxbury Press, Belmont. (1996).
2. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 2. John Wiley & Sons, New York. (1971).
3. James, B. R.: probabilidade: um curso em nível intermediario. IMPA, Rio de Janeiro. (2002).
4. Meester, R.: A Natural Introduction to Probability Theory. Birkhauser, Berlin. (2008).
5. Meyer, P. L.: Introductory Probability and Statistical Applications. Addison-Wesley, Massachusetts. (1972).
6. Ross, S.: Introduction to Probability Models. Academic Press, San Diego. (2007)
7. Soong, T. T.: Fundamentals of Probability and Statistics for Engineers. John Wiley & Sons Ltd. (2004).

Condicionales

(Borradores, Curso 23)

Sebastian Grynberg

8-10 de abril 2013



Serpientes de Akiyoshi Kitaoka.

*Si no se espera,
no se encontrará lo inesperado,
pues el sendero que a ello conduce
es inaccesible*
(Heráclito.)

Índice

1. Condicionales	2
1.1. Caso discreto	2
1.2. Mezclas	4
1.3. Sobre la regla de Bayes	5
1.4. Caso continuo	7
2. Predicción y Esperanza condicional	10
2.1. Ejemplos	12
2.1.1. Caso continuo	12
2.1.2. Regla de Bayes para mezclas	12
2.1.3. Caso discreto	13
2.2. Propiedades	14
2.3. Ejemplo: sumas aleatorias de variables aleatorias	16
2.4. Ejemplo: esperanza y varianza de una mezcla.	17
3. Predicción lineal y coeficiente de correlación	18
4. Bibliografía consultada	20

1. Condicionales

1.1. Caso discreto

Sean X e Y dos variables aleatorias discretas definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Fijemos un valor $x \in \mathbb{R}$ tal que $p_X(x) > 0$. Usando la noción de probabilidad condicional podemos definir la *función de probabilidad condicional de Y dado que $X = x$* , mediante

$$p_{Y|X=x}(y) := \mathbb{P}(Y = y | X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} = \frac{p_{X,Y}(x, y)}{p_X(x)}. \quad (1)$$

Función de distribución condicional de Y dado que $X = x$. La *función de distribución condicional de Y dado que $X = x$* se define por

$$F_{Y|X=x}(y) := \mathbb{P}(Y \leq y | X = x) = \sum_{z \leq y} \mathbb{P}(Y = z | X = x) = \sum_{z \leq y} p_{Y|X=x}(z). \quad (2)$$

Esperanza condicional de Y dado que $X = x$. La *esperanza condicional de Y dado que $X = x$* se define por

$$\mathbb{E}[Y | X = x] := \sum_y y p_{Y|X=x}(y). \quad (3)$$

Nota Bene 1. La función $F_{Y|X=x} : \mathbb{R} \rightarrow \mathbb{R}$ definida en (2) es una *función de distribución genuina*: es no decreciente, continua a derecha, tiende a 0 cuando $y \rightarrow -\infty$ y tiende a 1 cuando $y \rightarrow \infty$. Por lo tanto, podemos interpretarla como la función de distribución de una nueva variable aleatoria, $Y|X = x$, cuya ley de distribución coincide con la de Y cuando se sabe que ocurrió el evento $X = x$. Motivo por el cual la llamaremos *Y condicional a que $X = x$* .

Nota Bene 2. Todas las nociones asociadas a las distribuciones condicionales se definen de la misma manera que en el caso de una única variable aleatoria discreta, salvo que ahora todas las probabilidades se determinan condicionales al evento $X = x$. Las definiciones tienen sentido siempre y cuando $x \in Sop(p_X)$.

Nota Bene 3. Si se quieren calcular las funciones de probabilidad de las variables $Y|X = x$, $x \in Sop(p_X)$, la fórmula (1) dice que basta dividir cada fila de la representación matricial de la función de probabilidad conjunta de X e Y , $p_{X,Y}(x,y)$ por el correspondiente valor de su margen derecho, $p_X(x)$. En la fila x de la matriz resultante se encuentra la función de probabilidad condicional de Y dado que $X = x$, $p_{Y|X=x}(y)$.

Ejemplo 1.1. En una urna hay 3 bolas rojas, 2 amarillas y 1 verde. Se extraen dos. Sean X e Y la cantidad de bolas rojas y amarillas extraídas, respectivamente. La representación matricial de la función de probabilidad conjunta $p_{X,Y}(x,y)$ y de sus marginales $p_X(x)$, $p_Y(y)$ es la siguiente

$X \setminus Y$	0	1	2	p_X
0	0	2/15	1/15	3/15
1	3/15	6/15	0	9/15
2	3/15	0	0	3/15
p_Y	6/15	8/15	1/15	

Cuadro 1: Distribución conjunta de X e Y y sus respectivas marginales.

Dividiendo cada fila de la matriz $p_{X,Y}(x,y)$ por el correspondiente valor de su margen derecho se obtiene el Cuadro 2 que contiene toda la información sobre las funciones de probabilidad de las condicionales $Y|X = x$.

$X \setminus Y$	0	1	2
0	0	2/3	1/3
1	1/3	2/3	0
2	1	0	0

Cuadro 2: Distribuciones de las variables condicionales Y dado que $X = x$. Interpretación intuitiva de los resultados: *a medida que X aumenta el grado de indeterminación de Y disminuye*.

Por ejemplo, la función de probabilidad condicional de Y dado que $X = 0$, es la función de y definida en la primera fila del Cuadro 2: $p_{Y|X=0}(0) = 0$, $p_{Y|X=0}(1) = 2/3$ y $p_{Y|X=0}(2) = 1/3$.

Notar que la función de probabilidad condicional obtenida es diferente de la correspondiente a la marginal de Y , $p_Y(y)$. Del Cuadro 2 y la definición (3) se deduce que

$$\mathbb{E}[Y|X = x] = \frac{4}{3}\mathbf{1}\{x = 0\} + \frac{2}{3}\mathbf{1}\{x = 1\}. \quad (4)$$

□

Nota Bene. Observar que en general la función de probabilidad condicional $p_{Y|X=x}(y)$ es diferente de la función de probabilidad $p_Y(y)$. Esto indica que se pueden hacer inferencias sobre los valores posibles de Y a partir de los valores observados de X y viceversa; las dos variables son (estocásticamente) *dependientes*. Más adelante veremos algunas maneras de hacer este tipo de inferencias.

1.2. Mezclas

Definición 1.2 (Mezcla). Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad. Sea $M : \Omega \rightarrow \mathbb{R}$ una variable aleatoria discreta tal que $M(\Omega) = \mathcal{M}$ y $p_M(m) = \mathbb{P}(M = m) > 0$ para todo $m \in \mathcal{M}$. Sea $(X_m : m \in \mathcal{M})$ una familia de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ e independiente de M . En tal caso, la variable aleatoria $X := X_M$ está bien definida y se llama la *mezcla* de las variables X_m obtenida mediante la variable *mezcladora* M .

Nota Bene. La distribución de probabilidades de M indica la proporción en que deben mezclarse las variables X_m : para cada $m \in \mathcal{M}$, la probabilidad $p_M(m)$ representa la proporción con que la variable X_m participa de la mezcla X_M .

Cálculo de la función de distribución. La función de distribución de la mezcla X se obtiene utilizando la fórmula de probabilidad total:

$$\begin{aligned} F_X(x) &= \mathbb{P}(X_M \leq x) = \sum_{m \in \mathcal{M}} \mathbb{P}(X_M \leq x | M = m) \mathbb{P}(M = m) \\ &= \sum_{m \in \mathcal{M}} \mathbb{P}(X_m \leq x | M = m) p_M(m) \\ &= \sum_{m \in \mathcal{M}} \mathbb{P}(X_m \leq x) p_M(m) \quad (\text{pues } (X_m : m \in \mathcal{M}) \text{ y } M \text{ son indep.}) \\ &= \sum_{m \in \mathcal{M}} F_{X_m}(x) p_M(m), \end{aligned} \quad (5)$$

donde, para cada $m \in \mathcal{M}$, $F_{X_m}(x) = \mathbb{P}(X_m \leq x)$ es la función de distribución de la variable X_m .

Variables discretas. Si las variables aleatorias X_m son discretas con funciones de probabilidad $p_{X_m}(x) = \mathbb{P}(X_m = x)$, respectivamente, la mezcla X es discreta y su función de probabilidad es

$$p_X(x) = \sum_{m \in \mathcal{M}} p_{X_m}(x) p_M(m). \quad (6)$$

Variables absolutamente continuas. Si las variables X_m son absolutamente continuas con densidades $f_{X_m}(x)$, respectivamente, la mezcla X es absolutamente continua y tiene densidad

$$f_X(x) = \sum_{m \in \mathcal{M}} f_{X_m}(x)p_M(m). \quad (7)$$

Ejemplo 1.3. Para simular los valores de una variable aleatoria X se recurre al siguiente algoritmo: se simula el valor de un variable aleatoria M con distribución Bernoulli de parámetro $p = 1/5$. Si $M = 0$, se simula el valor de una variable aleatoria X_0 con distribución uniforme sobre el intervalo $(0, 4)$. Si $M = 1$, se simula el valor de una variable aleatoria X_1 con distribución uniforme sobre el intervalo $(2, 6)$. Se quiere hallar la densidad de probabilidades de la variable X así simulada.

La variable X es una mezcla. La variable mezcladora es M y las variables aleatorias que componen la mezcla son X_0 y X_1 . Por hipótesis, la variable mezcladora M se distribuye de acuerdo con la función de probabilidad $p_M(0) = 4/5$, $p_M(1) = 1/5$ y las distribuciones de las variables componentes son $X_0 \sim \mathcal{U}(0, 4)$ y $X_1 \sim \mathcal{U}(2, 6)$. En otras palabras, las densidades de las variables componente son $f_{X_0}(x) = \frac{1}{4}\mathbf{1}\{0 < x < 4\}$ y $f_{X_1}(x) = \frac{1}{4}\mathbf{1}\{2 < x < 6\}$. Usando la fórmula de probabilidad total (7) se obtiene la densidad de la mezcla X

$$\begin{aligned} f_X(x) &= p_M(0)f_{X_0}(x) + p_M(1)f_{X_1}(x) = \left(\frac{4}{5}\right)\frac{1}{4}\mathbf{1}\{0 < x < 4\} + \left(\frac{1}{5}\right)\frac{1}{4}\mathbf{1}\{2 < x < 6\} \\ &= \frac{4}{20}\mathbf{1}\{0 < x \leq 2\} + \frac{5}{20}\mathbf{1}\{2 < x < 4\} + \frac{1}{20}\mathbf{1}\{4 \leq x < 6\}. \end{aligned} \quad (8)$$

□

1.3. Sobre la regla de Bayes

Sean $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad; $M : \Omega \rightarrow \mathbb{R}$ una variable aleatoria discreta tal que $M(\Omega) = \mathcal{M}$ y $p_M(m) = \mathbb{P}(M = m) > 0$ para todo $m \in \mathcal{M}$. Sea $(X_m : m \in \mathcal{M})$ una familia de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ e independiente de M . Supongamos además que las variables $X_m, m \in \mathcal{M}$ son absolutamente continuas con densidades de probabilidad continuas $f_{X_m}(x), m \in \mathcal{M}$, respectivamente.

Sea $X := X_M$ la mezcla de las variables M_m obtenida mediante la variable mezcladora M . ¿Qué sentido debería tener la expresión $\mathbb{P}(M = m | X = x)$? No debe olvidarse que la variable X es absolutamente continua y en consecuencia $\mathbb{P}(X = x) = 0$. Por lo tanto, no tiene ningún sentido definir $\mathbb{P}(M = m | X = x)$ mediante un cociente de la forma

$$\mathbb{P}(M = m | X = x) = \frac{\mathbb{P}(X = x, M = m)}{\mathbb{P}(X = x)} = \frac{0}{0}.$$

¿Qué hacer? El obstáculo se puede superar siempre y cuando $f_X(x) > 0$. En tal caso, si “engordamos” el punto x mediante el intervalo de radio $h > 0$ (suficientemente chico) centrado en x , $B_h(x) := \{x - h < t < x + h\}$, el evento $\{X \in B_h(x)\}$ tiene probabilidad positiva

$$\mathbb{P}(X \in B_h(x)) = \int_{x-h}^{x+h} f_Y(t)dt = 2hf_X(\theta(h)), \quad \theta(h) \in B_h(x). \quad (9)$$

y la probabilidad condicional del evento $\{M = m\}$, dado que ocurrió el evento $\{X \in B_h(x)\}$ está bien definida y vale

$$\mathbb{P}(M = m | X \in B_h(x)) = \frac{\mathbb{P}(M = m, X \in B_h(x))}{\mathbb{P}(X \in B_h(x))}.$$

Por otra parte,

$$\begin{aligned}\mathbb{P}(M = m, X \in B_h(x)) &= p_M(m) \mathbb{P}(X_m \in B_h(x) | M = m) = p_M(m) \mathbb{P}(X_m \in B_h(x)) \\ &= p_M(m) \int_{x-h}^{x+h} f_{X_m}(t) dt = 2hp_M(m) f_{X_m}(\theta_m(h)),\end{aligned}\quad (10)$$

para algún $\theta_m(h) \in B_h(x)$. De (9) y (10) se deduce que

$$\mathbb{P}(M = m | X \in B_h(x)) = \frac{p_M(m) f_{X_m}(\theta_m(h))}{f_X(\theta(h))} \quad (11)$$

Para “adelgazar” el punto “engordado” hacemos $h \rightarrow 0$ y obtenemos

$$\lim_{h \rightarrow 0} \mathbb{P}(M = m | X \in B_h(x)) = \lim_{h \rightarrow 0} \frac{p_M(m) f_{X_m}(\theta_m(h))}{f_X(\theta(h))} = \frac{p_M(m) f_{X_m}(x)}{f_X(x)}. \quad (12)$$

Finalmente, para cada $x \in \mathbb{R}$ tal que $f_X(x) > 0$ definimos $\mathbb{P}(M = m | X = x)$ mediante la fórmula

$$\mathbb{P}(M = m | X = x) := \frac{p_M(m) f_{X_m}(x)}{f_X(x)}. \quad (13)$$

Ejemplo 1.4 (Detección de señales). Un emisor transmite un mensaje binario en la forma de una señal aleatoria Y que puede ser -1 o $+1$ con igual probabilidad. El canal de comunicación corrompe la transmisión con un ruido normal aditivo de media 0 y varianza 1 . El receptor recibe la señal $X = N + Y$, donde N es un ruido (*noise*) con distribución $\mathcal{N}(0, 1)$, independiente de Y . La pregunta del receptor es la siguiente: dado que recibí el valor x , cuál es la probabilidad de que la señal sea 1 ?

La señal que recibe el receptor es una mezcla. La variable mezcladora es Y y las variables aleatorias que componen la mezcla son $X_{-1} = N - 1$ y $X_1 = N + 1$. Por hipótesis, la variable mezcladora Y se distribuye de acuerdo con la función de probabilidad $p_Y(-1) = p_Y(1) = 1/2$ y las distribuciones de las variables componentes son $X_{-1} \sim \mathcal{N}(-1, 1)$ y $X_1 \sim \mathcal{N}(1, 1)$. En otras palabras, las densidades de las variables componente son

$$f_{X_{-1}}(x) = \frac{1}{\sqrt{2\pi}} e^{-(x+1)^2/2} \quad \text{y} \quad f_{X_1}(x) = \frac{1}{\sqrt{2\pi}} e^{-(z-1)^2/2}.$$

Usando la fórmula de probabilidad total (7) se obtiene la densidad de la mezcla X

$$f_X(x) = p_Y(-1)f_{X_{-1}}(x) + p_Y(1)f_{X_1}(x) = \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}} e^{-(x+1)^2/2} \right) + \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}} e^{-(z-1)^2/2} \right).$$

El receptor pregunta $\mathbb{P}(Y = 1 | X = x) = ?$ La respuesta se obtiene usando la regla de Bayes (13)

$$\mathbb{P}(Y = 1 | X = x) = \frac{p_Y(1)f_{X_1}(x)}{f_X(x)} = \frac{e^{-(x-1)^2/2}}{e^{-(x-1)^2/2} + e^{-(x+1)^2/2}} = \frac{e^x}{e^x + e^{-x}}. \quad (14)$$

□

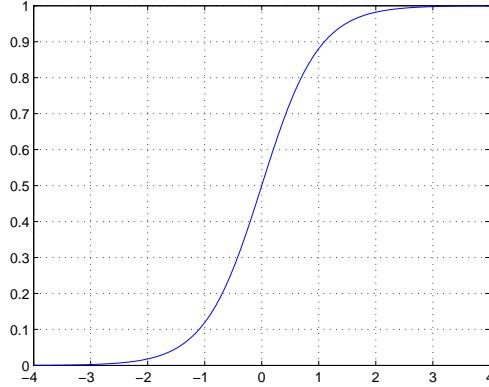


Figura 1: Gráfico de la probabilidad condicional $\mathbb{P}(Y = 1 | X = \cdot) : \mathbb{R} \rightarrow \mathbb{R}$ vista como función de x .

1.4. Caso continuo

Sean X e Y dos variables aleatorias definidas sobre $(\Omega, \mathcal{A}, \mathbb{P})$ con densidad conjunta $f_{X,Y}(x, y)$ continua. A diferencia del caso en que X es discreta en este caso tenemos que $\mathbb{P}(X = x) = 0$ para todo $x \in \mathbb{R}$, lo que hace imposible definir la función de distribución condicional de Y dado que $X = x$, $\mathbb{P}(Y \leq y | X = x)$, mediante el cociente (2):

$$\frac{\mathbb{P}(Y \leq y, X = x)}{\mathbb{P}(X = x)} = \frac{0}{0}.$$

Este obstáculo se puede superar observando que para cada $x \in Sop(f_X)$ y para cada $h > 0$ el evento $\{X \in B_h(x)\} = \{x - h < X < x + h\}$ tiene probabilidad positiva

$$\mathbb{P}(X \in B_h(x)) = \int_{x-h}^{x+h} f_X(s) ds = 2h f_X(\theta_1(h)), \quad \theta_1(h) \in B_h(x).$$

Por otra parte,

$$\mathbb{P}(Y \leq y, X \in B_h(x)) = \int_{x-h}^{x+h} \left(\int_{-\infty}^y f_{X,Y}(s, t) dt \right) ds = 2h \int_{-\infty}^y f_{X,Y}(\theta_2(h), t) dt,$$

donde $\theta_2(h) \in B_h(x)$.

Si $x \in Sop(f_X)$, la probabilidad condicional $\mathbb{P}(Y \leq y | X \in B_h(x))$ está bien definida y vale

$$\mathbb{P}(Y \leq y | X \in B_h(x)) = \frac{\mathbb{P}(Y \leq y, X \in B_h(x))}{\mathbb{P}(X \in B_h(x))} = \frac{\int_{-\infty}^y f_{X,Y}(\theta_2(h), t) dt}{f_X(\theta_1(h))}.$$

En consecuencia,

$$\lim_{h \rightarrow 0} \mathbb{P}(Y \leq y | X \in B_h(x)) = \frac{\int_{-\infty}^y f_{X,Y}(x, t) dt}{f_X(x)}. \quad (15)$$

El lado derecho de (15) define una genuina función de distribución $F_{Y|X=x} : \mathbb{R} \rightarrow \mathbb{R}$,

$$F_{Y|X=x}(y) := \frac{\int_{-\infty}^y f_{X,Y}(x,t) dt}{f_X(x)}, \quad (16)$$

que se llama la *función distribución condicional de Y dado X = x* y se puede interpretar como la función de distribución de una nueva variable aleatoria que llamaremos *Y condicional a que X = x* y que será designada mediante el símbolo $Y|X = x$.

La función de distribución $F_{Y|X=x}(y)$ es derivable y su derivada

$$f_{Y|X=x}(y) := \frac{d}{dy} F_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad (17)$$

se llama la *densidad condicional de Y dado que X = x*.

Curva peligrosa. Todo el argumento usa la hipótesis $f_X(x) > 0$. Si $f_X(x) = 0$ las expresiones (15)-(17) carecen de sentido. Sin embargo, esto no es un problema grave ya que $\mathbb{P}(X \in Sop(f_X)) = 1$. Para los valores de x tales que $f_X(x) = 0$ las variables condicionales $Y|X = x$ serán definidas como idénticamente nulas. En tal caso, $F_{Y|X=x}(y) = \mathbf{1}\{y \geq 0\}$.

Regla mnemotécnica. De la fórmula (17) se deduce que $f_{X,Y}(x,y) = f_{Y|X=x}(y)f_X(x)$ y puede recordarse mediante el siguiente “versito”: “*la densidad conjunta es igual a la densidad condicional por la marginal de la condición*”.

Ejemplo 1.5 (Dos etapas: conjunta = marginal \times condicional). Se elige un número al azar X sobre el intervalo $(0, 1)$ y después otro número al azar Y sobre el intervalo $(X, 1)$. Se quiere hallar la densidad marginal de Y . Por hipótesis, $f_X(x) = \mathbf{1}\{0 < x < 1\}$ y $f_{Y|X=x}(y) = \frac{1}{1-x}\mathbf{1}\{x < y < 1\}$. La densidad conjunta de X e Y se obtiene multiplicando la densidad condicional $f_{Y|X=x}(y)$ por la densidad marginal $f_X(x)$: $f_{X,Y}(x,y) = f_{Y|X=x}(y)f_X(x) = \frac{1}{1-x}\mathbf{1}\{0 < x < y < 1\}$. La densidad marginal de Y se obtiene integrando la densidad conjunta $f_{X,Y}(x,y)$ con respecto a x

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} \frac{1}{1-x} \mathbf{1}\{0 < x < y < 1\} dx = \mathbf{1}\{0 < y < 1\} \int_0^y \frac{1}{1-x} dx \\ &= -\log(1-y)\mathbf{1}\{0 < y < 1\}. \end{aligned}$$

□

Fórmula de probabilidad total. La densidad de probabilidades de Y es una combinación convexa de las condicionales:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X=x}(y)f_X(x) dx.$$

Inmediato de la relación “conjunta = marginal \times condicional”. Integrando respecto de y se obtiene que la función de distribución de Y es una combinación convexa de las condicionales:

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^y f_Y(t) dt = \int_{-\infty}^y \left(\int_{-\infty}^{\infty} f_{Y|X=x}(t)f_X(x) dx \right) dt \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^y f_{Y|X=x}(t) dt \right) f_X(x) dx = \int_{-\infty}^{\infty} F_{Y|X=x}(y)f_X(x) dx. \end{aligned}$$

Esperanza condicional de Y dado que $X = x$. Para cada $x \in \mathbb{R}$, la *esperanza condicional de Y dado que $X = x$* se define por

$$\mathbb{E}[Y|X = x] := \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy. \quad (18)$$

siempre y cuando la integral del converja absolutamente. Si $f_X(x) = 0$, $\mathbb{E}[Y|X = x] = 0$.

Varianza condicional

En cualquier caso, definidas las esperanzas condicionales de Y y de Y^2 dado que $X = x$, la *varianza condicional de Y dado que $X = x$* se define mediante

$$\mathbb{V}(Y|X = x) := \mathbb{E}[(Y - \mathbb{E}[Y|X = x])^2 | X = x] \quad (19)$$

Desarrollando el término derecho se obtiene

$$\mathbb{V}(Y|X = x) = \mathbb{E}[Y^2|X = x] - \mathbb{E}[Y|X = x]^2. \quad (20)$$

Nota Bene. La definición es consistente y coincide con la varianza de la variable aleatoria $Y|X = x$ cuya función de distribución es $F_{Y|X=x}(y)$.

Ejemplo 1.6 (Dardos). Volvamos al problema del juego de dardos de blanco circular $\Lambda = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$. Por hipótesis, el dardo se clava en un punto de coordenadas (X, Y) uniformemente distribuido sobre Λ .

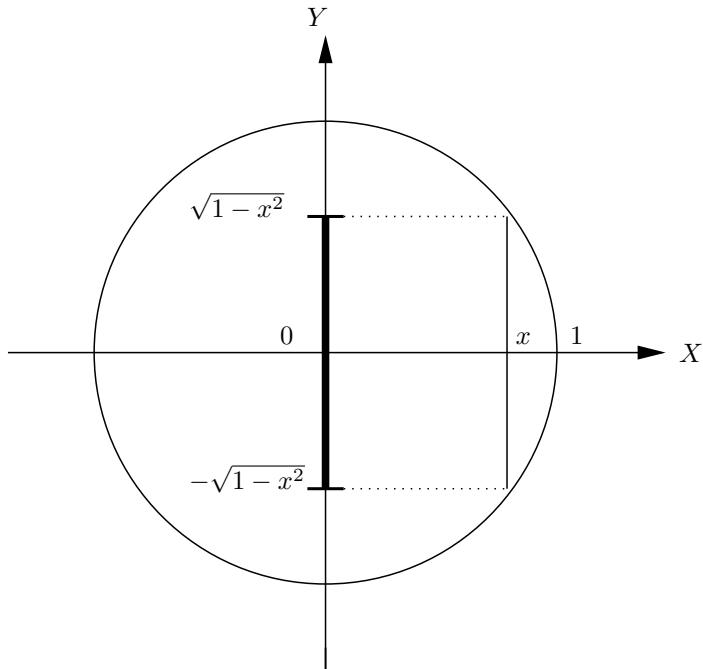


Figura 2: Para cada $x \in [-1, 1]$ se observa que $Y|X = x \sim \mathcal{U}[-\sqrt{1-x^2}, \sqrt{1-x^2}]$.

La densidad conjunta de X e Y es $f_{X,Y}(x,y) = \frac{1}{\pi} \mathbf{1}\{x^2 + y^2 \leq 1\}$. Por definición, para cada $x \in [-1, 1]$, la densidad condicional de Y dado que $X = x$ es el cociente entre la densidad conjunta $f_{X,Y}(x,y)$ y la densidad marginal de X

$$f_X(x) = \frac{2\sqrt{1-x^2}}{\pi} \mathbf{1}\{x \in [-1, 1]\}.$$

Por lo tanto,

$$f_{Y|X=x}(y) = \frac{1}{2\sqrt{1-x^2}} \mathbf{1}\{-\sqrt{1-x^2} \leq y \leq \sqrt{1-x^2}\}. \quad (21)$$

En otras palabras, dado que $X = x$, $x \in [-1, 1]$, la variable Y se distribuye uniformemente sobre el intervalo $[-\sqrt{1-x^2}, \sqrt{1-x^2}]$. En consecuencia,

$$\mathbb{E}[Y|X=x] = 0 \quad \text{y} \quad \mathbb{V}(Y|X=x) = (2\sqrt{1-x^2})^2/12 = (1-x^2)/3.$$

□

2. Predicción y Esperanza condicional

Planteo del problema

En su versión más simple un problema de predicción o estimación involucra dos variables aleatorias: una variable aleatoria Y desconocida (o inobservable) y una variable aleatoria X conocida (u observable). El problema consiste en deducir información sobre el valor de Y a partir del conocimiento del valor de X . Para ser más precisos, se busca una función $\varphi(X)$ que (en algún sentido) sea lo más parecida a Y como sea posible. La variable aleatoria $\hat{Y} := \varphi(X)$ se denomina un *estimador* de Y .

Ejemplo 2.1 (Detección de señales). Un emisor transmite un mensaje binario en la forma de una señal aleatoria Y que puede ser -1 o $+1$ con igual probabilidad. El canal de comunicación corrompe la transmisión con un ruido normal aditivo de media 0 y varianza σ^2 . El receptor recibe la señal $X = Y + N$, donde N es un ruido con distribución $\mathcal{N}(0, \sigma^2)$, independiente de Y . El receptor del mensaje observa la señal corrompida X y sobre esa base tiene que “reconstruir” la señal original Y . ¿Cómo lo hace?, ¿Qué puede hacer?

En lo que sigue desarrollaremos herramientas que permitan resolver este tipo de problemas. Sean X e Y dos variables aleatorias definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. El objetivo es construir una función $\varphi(X)$ que *sea lo más parecida a Y como sea posible*. En primer lugar, vamos a suponer que $\mathbb{E}[|Y|] < \infty$. Esta hipótesis permite precisar el sentido del enunciado *parecerse a Y* . Concretamente, queremos construir una función de X , $\varphi(X)$, que solucione la siguiente ecuación funcional

$$\mathbb{E}[\varphi(X)h(X)] = \mathbb{E}[Yh(X)], \quad (22)$$

para toda función medible y acotada $h : \mathbb{R} \rightarrow \mathbb{R}$.

Esperanza condicional

Sean X e Y dos variables aleatorias definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Supongamos que $\mathbb{E}[|Y|] < \infty$. Definimos la *esperanza condicional de Y dada X* , $\mathbb{E}[Y|X]$, como cualquier variable aleatoria de la forma $\varphi(X)$, donde $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ es una función (medible), que solucione la ecuación funcional (22).

Existencia. La existencia de la esperanza condicional depende de teoremas profundos de Teoría de la medida y no será discutida en estas notas. El lector interesado puede consultar Billingsley(1986) y/o Durrett(1996).

Unicidad. Supongamos que $\varphi(X)$ y $\psi(X)$ son dos soluciones de la ecuación funcional (22). Entonces, $\varphi(X) = \psi(X)$ casi seguramente (i.e., $\mathbb{P}(\varphi(X) \neq \psi(X)) = 0$).

Demostración. Por cuestiones de simetría, la prueba se reduce a mostrar que para cada $\epsilon > 0$, $\mathbb{P}(A_\epsilon) = 0$, donde $A_\epsilon := \{\varphi(X) - \psi(X) \geq \epsilon\}$. Observar que, por hipótesis, para toda función medible y acotada $h : \mathbb{R} \rightarrow \mathbb{R}$ vale que $\mathbb{E}[\varphi(X)h(X)] = \mathbb{E}[\psi(X)h(X)]$ o lo que es equivalente $\mathbb{E}[(\varphi(X) - \psi(X))h(X)] = 0$. Poniendo $h(X) = \mathbf{1}\{X \in A_\epsilon\}$ tenemos que $0 = \mathbb{E}[(\varphi(X) - \psi(X))\mathbf{1}\{X \in A_\epsilon\}] \geq \mathbb{E}[\epsilon\mathbf{1}\{X \in A_\epsilon\}] = \epsilon\mathbb{P}(A_\epsilon)$. Por lo tanto, $\mathbb{P}(A_\epsilon) = 0$. \square

Lema 2.2 (Técnico). La esperanza condicional satisface $\mathbb{E}[|\mathbb{E}[Y|X]|] \leq \mathbb{E}[|Y|]$.

Demostración. La variable aleatoria $\varphi(X)$ satisface la ecuación (22). Poniendo $h(X) = \mathbf{1}\{\varphi(X) > 0\}$ y usando (22) se obtiene

$$\mathbb{E}[\varphi(X)\mathbf{1}\{\varphi(X) > 0\}] = \mathbb{E}[Y\mathbf{1}\{\varphi(X) > 0\}] \leq \mathbb{E}[|Y|].$$

Análogamente se puede ver que $\mathbb{E}[-\varphi(X)\mathbf{1}\{\varphi(X) \leq 0\}] = \mathbb{E}[-Y\mathbf{1}\{\varphi(X) \leq 0\}] \leq \mathbb{E}[|Y|]$. Por lo tanto,

$$\begin{aligned} \mathbb{E}[|\varphi(X)|] &= \mathbb{E}[\varphi(X)\mathbf{1}\{\varphi(X) > 0\} - \varphi(X)\mathbf{1}\{\varphi(X) \leq 0\}] \\ &= \mathbb{E}[\varphi(X)\mathbf{1}\{\varphi(X) > 0\}] - \mathbb{E}[\varphi(X)\mathbf{1}\{\varphi(X) \leq 0\}] \\ &= \mathbb{E}[Y\mathbf{1}\{\varphi(X) > 0\}] - \mathbb{E}[Y\mathbf{1}\{\varphi(X) \leq 0\}] \\ &= \mathbb{E}[Y\mathbf{1}\{\varphi(X) > 0\} - Y\mathbf{1}\{\varphi(X) \leq 0\}] \leq \mathbb{E}[|Y|]. \end{aligned}$$

\square

Propiedades que merecen ser subrayadas

Aunque se deducen inmediatamente de la definición, las propiedades siguientes merecen ser subrayas porque, como se podrá apreciar más adelante, constituyen poderosas herramientas de cálculo.

1. Fórmula de probabilidad total:

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]. \quad (23)$$

2. Sea $g : \mathbb{R} \rightarrow \mathbb{R}$ una función tal que $\mathbb{E}[|g(X)Y|] < \infty$,

$$\mathbb{E}[g(X)Y|X] = g(X)\mathbb{E}[Y|X]. \quad (24)$$

3. Si X e Y son independientes, entonces $\mathbb{E}[Y|X] = \mathbb{E}[Y]$.

Demostración. La fórmula de probabilidad total se deduce de la ecuación (22) poniendo $h(X) \equiv 1$. La identidad (24) se obtiene observando que $g(X)\mathbb{E}[Y|X]$ es una función de X que soluciona la ecuación $\mathbb{E}[g(X)\mathbb{E}[Y|X]h(X)] = \mathbb{E}[(g(X)Y)h(X)]$. Si X e Y son independientes $\mathbb{E}[Yh(X)] = \mathbb{E}[Y]\mathbb{E}[h(X)] = \mathbb{E}[\mathbb{E}[Y]h(X)]$. \square

2.1. Ejemplos

2.1.1. Caso continuo

Sean X e Y dos variables aleatorias continuas definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ con densidad de probabilidades conjunta $f_{X,Y}(x,y)$ y $\mathbb{E}[|Y|] < \infty$. La esperanza condicional de Y dada X es $\mathbb{E}[Y|X] = \varphi(X)$, donde $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ es la función de regresión de Y sobre X definida por

$$\varphi(x) := \mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy. \quad (25)$$

Demostración. Basta ver $\varphi(X)$ verifica la ecuación funcional (22) para cualquier función h medible y acotada.

$$\begin{aligned} \mathbb{E}[\varphi(X)h(X)] &= \int_{-\infty}^{\infty} \varphi(x)h(x)f_X(x)dx = \int_{-\infty}^{\infty} \mathbb{E}[Y|X=x]h(x)f_X(x)dx \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy \right) h(x)f_X(x)dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y h(x) f_{Y|X=x}(y) f_X(x) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y h(x) f_{X,Y}(x,y) dx dy = \mathbb{E}[Yh(X)]. \end{aligned}$$

\square

2.1.2. Regla de Bayes para mezclas

Volvamos el Ejemplo 2.1 la pregunta es ¿Qué puede hacer el receptor para “reconstruir” la señal original, Y , a partir de la señal corrompida X ? Lo “mejor” que puede hacer es estimar Y mediante la esperanza condicional $\mathbb{E}[Y|X]$. El receptor recibe la mezcla de dos variables aleatorias $X|Y = -1 \sim \mathcal{N}(-1, \sigma^2)$ e $X|Y = 1 \sim \mathcal{N}(1, \sigma^2)$, mezcladas en igual proporción: $p_Y(-1) = p_Y(1) = 1/2$. Las densidades de las componentes de la mezcla son

$$f_{X|Y=-1}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x+1)^2/2\sigma^2} \quad \text{y} \quad f_{X|Y=1}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-1)^2/2\sigma^2}.$$

De la fórmula de probabilidad total se deduce que la densidad de la mezcla X es

$$\begin{aligned} f_X(x) &= p_Y(-1)f_{X|Y=-1}(x) + p_Y(1)f_{X|Y=1}(x) \\ &= \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(x+1)^2/2\sigma^2} \right) + \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-1)^2/2\sigma^2} \right). \end{aligned} \quad (26)$$

Para construir la esperanza condicional $\mathbb{E}[Y|X]$ el receptor debe calcular la función de regresión $\varphi(x) = \mathbb{E}[Y|X = x] = p_Y(1)f_{X|Y=1}(x) - p_Y(-1)f_{X|Y=-1}(x)$. Que de acuerdo con la regla de Bayes para mezclas adopta la forma

$$\varphi(x) = \frac{p_Y(1)f_{X|Y=1}(x) - p_Y(-1)f_{X|Y=-1}(x)}{f_X(x)} = \frac{e^{x/\sigma^2} - e^{-x/\sigma^2}}{e^{x/\sigma^2} + e^{-x/\sigma^2}} = \tanh(x/\sigma^2). \quad (27)$$

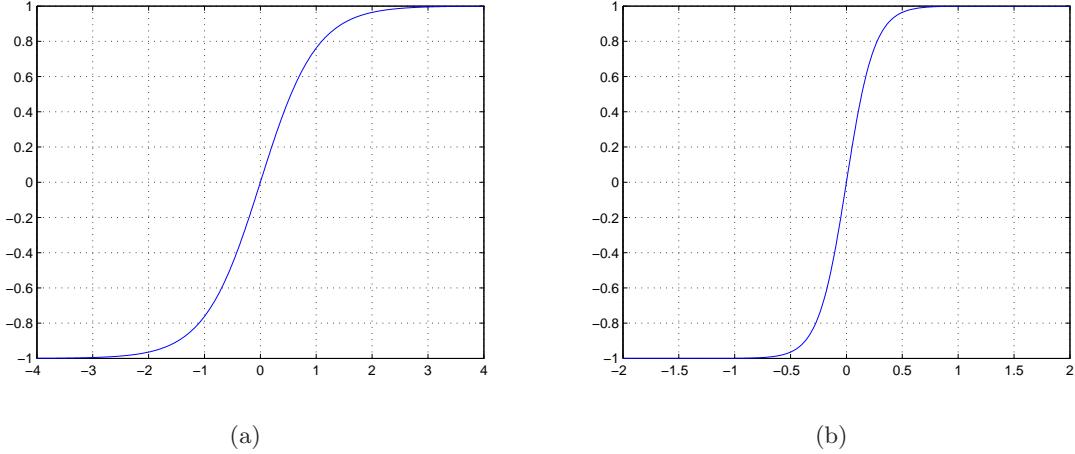


Figura 3: Líneas de regresión de Y sobre X para distintos valores de la varianza σ^2 . (a) $\sigma^2 = 1$: $\varphi(x) = \tanh(x)$; (b) $\sigma^2 = 1/4$, $\varphi(x) = \tanh(4x)$.

El receptor reconstruye Y basándose en X mediante $\mathbb{E}[Y|X] = \tanh(X/\sigma^2)$. □

2.1.3. Caso discreto

Sean X e Y dos variables aleatorias discretas definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, con función de probabilidad conjunta $p_{X,Y}(x, y)$ y $\mathbb{E}[|Y|] < \infty$. Para simplificar la exposición supongamos que $Sop(p_X) = X(\Omega)$. En tal caso, la esperanza condicional de Y dada X es $\mathbb{E}[Y|X] = \varphi(X)$, donde $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ es la función de regresión de Y sobre X definida por

$$\varphi(x) := \mathbb{E}[Y|X = x] = \sum_{y \in Y(\Omega)} y p_{Y|X=x}(y) \quad (28)$$

Demostración. Basta ver $\varphi(X)$ verifica la ecuación funcional (22) para cualquier función h medible y acotada.

$$\begin{aligned} \mathbb{E}[\varphi(X)h(X)] &= \sum_x \varphi(x)h(x)p_X(x) = \sum_x \mathbb{E}[Y|X = x]h(x)p_X(x) \\ &= \sum_x \left(\sum_y y p_{Y|X=x}(y) \right) h(x)p_X(x) = \sum_x \sum_y y h(x)p_{Y|X=x}(y)p_X(x) \\ &= \sum_x \sum_y y h(x)p_{X,Y}(x, y) = \mathbb{E}[Yh(X)]. \end{aligned}$$

□

Ejemplo 2.3 (Fórmula de probabilidad total). Una rata está atrapada en un laberinto. Inicialmente puede elegir una de tres direcciones. Si elige la primera se perderá en el laberinto y luego de 4 minutos volverá a su posición inicial; si elige la segunda volverá a su posición inicial luego de 7 minutos; si elige la tercera saldrá del laberinto luego de 3 minutos. Suponiendo que en cada intento, la rata elige con igual probabilidad cualquiera de las tres direcciones, cuál es la esperanza del tiempo que demora en salir del laberinto?

Sean Y la cantidad de tiempo que demora la rata en salir del laberinto y sea X la dirección que elige inicialmente. Usando la fórmula de probabilidad total puede verse que

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \sum_{x=1}^3 \mathbb{E}[Y|X=x]\mathbb{P}(X=x) = \frac{1}{3} \sum_{x=1}^3 \mathbb{E}[Y|X=x]$$

Si la rata elige la primera dirección, se pierde en el laberinto durante 4 minutos y vuelve a su posición inicial. Una vez que vuelve a su posición inicial el problema se renueva y la esperanza del tiempo adicional hasta que la rata consiga salir del laberinto es $\mathbb{E}[Y]$. En otros términos $\mathbb{E}[Y|X=1] = 4 + \mathbb{E}[Y]$. Análogamente puede verse que $\mathbb{E}[Y|X=2] = 7 + \mathbb{E}[Y]$. La igualdad $\mathbb{E}[Y|X=3] = 3$ no requiere comentarios. Por lo tanto,

$$\mathbb{E}[Y] = \frac{1}{3} (4 + \mathbb{E}[Y] + 7 + \mathbb{E}[Y] + 3) = \frac{1}{3} (2\mathbb{E}[Y] + 14).$$

Finalmente, $\mathbb{E}[Y] = 14$.

□

2.2. Propiedades

La esperanza condicional tiene propiedades similares a la esperanza.

Linealidad. $\mathbb{E}[aY_1 + bY_2|X] = a\mathbb{E}[Y_1|X] + b\mathbb{E}[Y_2|X]$.

Monotonía. Si $Y_1 \leq Y_2$, entonces $\mathbb{E}[Y_1|X] \leq \mathbb{E}[Y_2|X]$.

Desigualdad de Jensen. Si $g : \mathbb{R} \rightarrow \mathbb{R}$ es una función convexa y $\mathbb{E}[|Y|], \mathbb{E}[|g(Y)|] < \infty$, entonces

$$g(\mathbb{E}[Y|X]) \leq \mathbb{E}[g(Y)|X]. \quad (29)$$

En particular, si $\mathbb{E}[Y^2] < \infty$, poniendo $g(t) = t^2$ en la desigualdad de Jensen se obtiene

$$\mathbb{E}[Y|X]^2 \leq \mathbb{E}[Y^2|X] \quad (30)$$

Definición 2.4 (Varianza condicional). Sean X e Y dos variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Si $\mathbb{E}[Y^2] < \infty$, la *varianza condicional de Y dada X* , $\mathbb{V}(Y|X)$, se define por

$$\mathbb{V}(Y|X) := \mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2 \quad (31)$$

Predicción

Existen diversas maneras en las que dos variables pueden considerarse cercanas entre sí. Una manera es trabajar con la norma dada por $\|X\| := \sqrt{\mathbb{E}[X^2]}$ y definir la distancia entre dos variables aleatorias X e Y , $d(X, Y)$ mediante

$$d(X, Y) := \|Y - X\| = \sqrt{\mathbb{E}[(Y - X)^2]}. \quad (32)$$

Definición 2.5 (Predictor). Sean X e Y variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, tales que $\mathbb{E}[Y^2] < \infty$. El predictor de error cuadrático medio mínimo (o *mejor predictor*) de Y dada X es la función $\hat{Y} = h(X)$ de X que minimiza la distancia $d(\hat{Y}, Y)$ definida en (32).

El mejor predictor de Y dada X es una variable aleatoria \hat{Y} perteneciente al espacio vectorial $\mathbb{H} = \{h(X) : h : \mathbb{R} \rightarrow \mathbb{R}, \mathbb{E}[h(X)^2] < \infty\}$ tal que $\mathbb{E}[(Y - \hat{Y})^2] \leq \mathbb{E}[(Y - Z)^2]$ para toda $Z \in \mathbb{H}$.

Interpretación geométrica. Sea $L_2(\Omega, \mathcal{A}, \mathbb{P})$ el conjunto de todas las variables aleatorias definidas sobre $(\Omega, \mathcal{A}, \mathbb{P})$ que tienen varianza finita. \mathbb{H} es un subespacio de $L_2(\Omega, \mathcal{A}, \mathbb{P})$. Si $Y \notin \mathbb{H}$ entonces el camino más corto desde Y hasta \mathbb{H} es por la recta ortogonal al subespacio \mathbb{H} que pasa por Y . Por lo tanto, \hat{Y} debe ser la proyección ortogonal de Y sobre \mathbb{H} . En tal caso $Y - \hat{Y}$ es ortogonal a cualquier vector de \mathbb{H} . En otras palabras, $\langle Y - \hat{Y}, Z \rangle = 0$ para todo $Z \in \mathbb{H}$, donde $\langle X, Y \rangle$ es el producto interno en $L_2(\Omega, \mathcal{A}, \mathbb{P})$ definido por $\langle X, Y \rangle := \mathbb{E}[XY]$.

La esperanza condicional $\mathbb{E}[Y|X]$ es el mejor predictor de Y basado en X

- 1) La condición $\mathbb{E}[Y^2] < \infty$ implica que $\mathbb{E}[Y|X] \in \mathbb{H}$:

$$\mathbb{E}[\mathbb{E}[Y|X]^2] \leq \mathbb{E}[\mathbb{E}[Y^2|X]] = \mathbb{E}[Y^2] < \infty.$$

- 2) La ecuación funcional (22) significa que $Y - \mathbb{E}[Y|X] \perp \mathbb{H}$:

$$\begin{aligned} \langle Y - \mathbb{E}[Y|X], h(X) \rangle &= 0 \iff \mathbb{E}[(Y - \mathbb{E}[Y|X])h(X)] = 0 \\ &\iff \mathbb{E}[\mathbb{E}[Y|X]h(X)] = \mathbb{E}[Yh(X)]. \end{aligned}$$

Por lo tanto, la esperanza condicional, $\mathbb{E}[Y|X]$, satisface las dos condiciones que caracterizan a la proyección ortogonal sobre el subespacio \mathbb{H} y en consecuencia es el predictor de Y basado en X de menor error cuadrático:

$$\mathbb{E}[Y|X] = \arg \min_{h(X) \in \mathbb{H}} \mathbb{E}[(Y - h(X))^2].$$

El error cuadrático medio mínimo se puede expresar en la forma

$$\begin{aligned} \|Y - \mathbb{E}[Y|X]\|^2 &= \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] = \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]] \\ &= \mathbb{E}[\mathbb{V}(Y|X)]. \end{aligned}$$

La última igualdad se obtiene desarrollando el cuadrado $(Y - \mathbb{E}[Y|X])^2$ y usando las propiedades de la esperanza condicional. (*Ejercicio*)

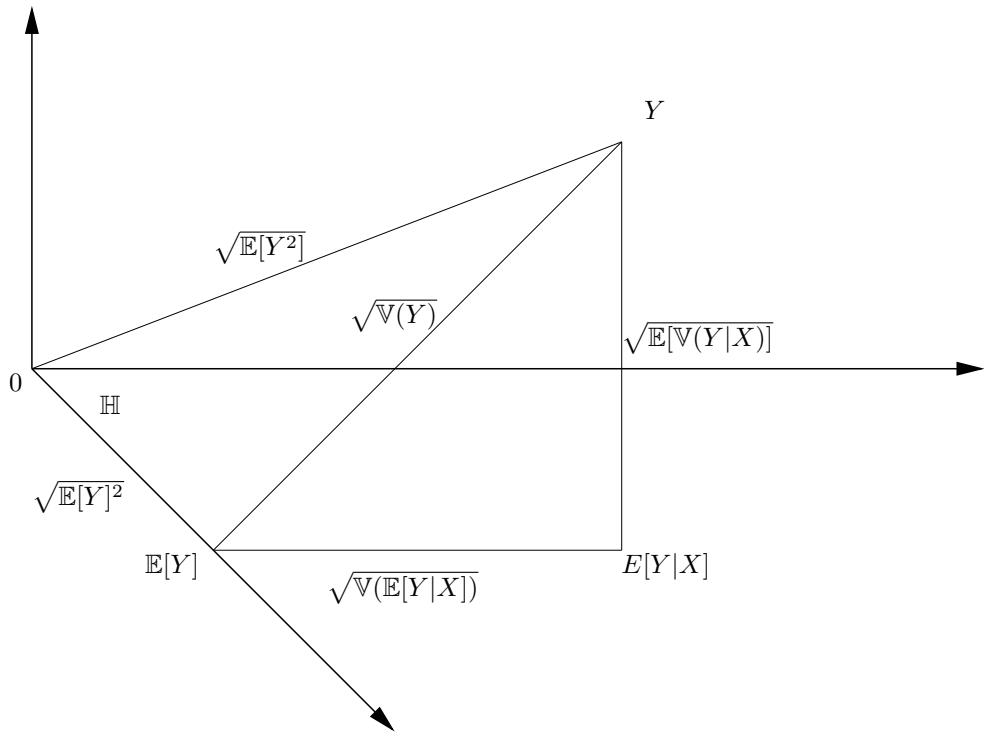


Figura 4: Teorema de Pitágoras: $\mathbb{V}(Y) = \mathbb{E}[\mathbb{V}(Y|X)] + \mathbb{V}(\mathbb{E}[Y|X])$.

Por último, como $\mathbb{E}[Y] \in \mathbb{H}$, el Teorema de Pitágoras implica que

$$\begin{aligned}\mathbb{V}(Y) &= \|Y - \mathbb{E}[Y]\|^2 = \|Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - \mathbb{E}[Y]\|^2 \\ &= \|Y - \mathbb{E}[Y|X]\|^2 + \|\mathbb{E}[Y|X] - \mathbb{E}[Y]\|^2 = \mathbb{E}[\mathbb{V}(Y|X)] + \mathbb{V}(\mathbb{E}[Y|X]).\end{aligned}\quad (33)$$

En otras palabras, la variabilidad de Y se descompone de la siguiente manera: la variabilidad (media) de Y alrededor de su esperanza condicional, más la variabilidad de esta última.

2.3. Ejemplo: sumas aleatorias de variables aleatorias

Sea X_1, X_2, \dots una sucesión de variables aleatorias idénticamente distribuidas de media μ y varianza σ^2 . Sea N una variable discreta a valores en \mathbb{N} que es independiente de las X_i . El problema consiste en hallar la media y la varianza de la variable aleatoria $S = \sum_{i=1}^N X_i$, llamada *variable aleatoria compuesta*. Este problema se puede resolver utilizando las identidades

$$\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S|N]] \quad \text{y} \quad \mathbb{V}(S) = \mathbb{E}[\mathbb{V}(S|N)] + \mathbb{V}(\mathbb{E}[S|N]).$$

En la jerga probabilística esta técnica de cálculo se conoce bajo el nombre de *cálculo de esperanzas y varianzas mediante condicionales*.

Cálculo de la esperanza por condicionales.

$$\begin{aligned}
\mathbb{E}[S|N=n] &= \mathbb{E}\left[\sum_{i=1}^N X_i \mid N=n\right] = \mathbb{E}\left[\sum_{i=1}^n X_i \mid N=n\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n X_i\right] \quad \text{por la independencia de las } X_i \text{ y } N \\
&= n\mu.
\end{aligned}$$

En consecuencia, $\mathbb{E}[S|N] = \mu N$. Por lo tanto, $\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S|N]] = \mathbb{E}[\mu N] = \mu \mathbb{E}[N]$. □

Cálculo de la varianza por condicionales.

$$\begin{aligned}
\mathbb{V}(S|N=n) &= \mathbb{V}\left(\sum_{i=1}^N X_i \mid N=n\right) = \mathbb{V}\left(\sum_{i=1}^n X_i \mid N=n\right) \\
&= \mathbb{V}\left(\sum_{i=1}^n X_i\right) \quad \text{por la independencia de } X_i \text{ y } N \\
&= n\sigma^2.
\end{aligned}$$

En consecuencia, $\mathbb{V}(S|N) = \sigma^2 N$. Por lo tanto, $\mathbb{E}[\mathbb{V}(S|N)] = \mathbb{E}[\sigma^2 N] = \sigma^2 \mathbb{E}[N]$. Por otra parte, $\mathbb{V}[\mathbb{E}(S|N)] = \mathbb{V}[\mu N] = \mu^2 \mathbb{V}[N]$. Finalmente,

$$\mathbb{V}(S) = \mathbb{E}[\mathbb{V}(S|N)] + \mathbb{V}(\mathbb{E}[S|N]) = \sigma^2 \mathbb{E}[N] + \mu^2 \mathbb{V}[N].$$

□

2.4. Ejemplo: esperanza y varianza de una mezcla.

Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad. Sea $M : \Omega \rightarrow \mathbb{R}$ una variable aleatoria discreta tal que $M(\Omega) = \mathcal{M}$ y $p_M(m) = \mathbb{P}(M=m) > 0$ para todo $m \in \mathcal{M}$ y sea $(X_m : m \in \mathcal{M})$ una familia de variables aleatorias definidas sobre el mismo espacio de probabilidad, independiente de M . El problema consiste en hallar la media y la varianza de la mezcla $X := X_M$.

La forma natural de resolver este problema es usar la técnica del *cálculo de esperanzas y varianzas mediante condicionales*:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|M]] \quad \text{y} \quad \mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X|M)] + \mathbb{V}(\mathbb{E}[X|M]).$$

Cálculo de la esperanza por condicionales. En primer lugar hay que observar que $X|M = m \sim X_m$ por lo tanto,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|M]] = \sum_{m \in \mathcal{M}} \mathbb{E}[X|M=m] \mathbb{P}(M=m) = \sum_{m \in \mathcal{M}} \mathbb{E}[X_m] p_M(m).$$

□

Cálculo de la varianza por condicionales.

$$\mathbb{E}[\mathbb{V}(X|M)] = \sum_{m \in \mathcal{M}} \mathbb{V}(X|M=m) \mathbb{P}(M=m) = \sum_{m \in \mathcal{M}} \mathbb{V}(X_m) p_M(m).$$

Por otra parte,

$$\begin{aligned} \mathbb{V}(\mathbb{E}[X|M]) &= \mathbb{E}[(\mathbb{E}[X|M] - \mathbb{E}[X])^2] = \sum_{m \in \mathcal{M}} (\mathbb{E}[X|M=m] - \mathbb{E}[X])^2 \mathbb{P}(M=m) \\ &= \sum_{m \in \mathcal{M}} (\mathbb{E}[X_m] - \mathbb{E}[X])^2 p_M(m). \end{aligned}$$

Finalmente,

$$\mathbb{V}(X) = \sum_{m \in \mathcal{M}} \mathbb{V}(X_m) p_M(m) + \sum_{m \in \mathcal{M}} (\mathbb{E}[X_m] - \mathbb{E}[X])^2 p_M(m).$$

Nota Bene. Comparar con el Teorema de Steiner para el momento de inercia. \square

3. Predicción lineal y coeficiente de correlación

Definición 3.1 (Predictor lineal). Sean X e Y dos variables aleatorias definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, tales que $\mathbb{E}[X^2] < \infty$ y $\mathbb{E}[Y^2] < \infty$. La *recta de regresión de Y basada en X* es la función lineal $\hat{Y} = aX + b$ que minimiza la distancia

$$d(\hat{Y}, Y) = \sqrt{\mathbb{E}[(Y - \hat{Y})^2]}.$$

Cálculo explícito de la recta de regresión. El problema consiste en hallar los valores de a y b que minimizan la siguiente función de dos variables

$$g(a, b) := \mathbb{E}[(Y - (aX + b))^2].$$

Usando técnicas de cálculo diferencial en varias variables el problema se reduce a resolver el sistema de ecuaciones $\nabla g = 0$. Desarrollando cuadrados se puede ver que

$$\begin{aligned} \frac{\partial g(a, b)}{\partial a} &= 2a\mathbb{E}[X^2] - 2\mathbb{E}[XY] + 2b\mathbb{E}[X], \\ \frac{\partial g(a, b)}{\partial b} &= 2b - 2\mathbb{E}[Y] + 2a\mathbb{E}[X]. \end{aligned}$$

El problema se reduce a resolver el siguiente sistema lineal de ecuaciones

$$\begin{cases} a\mathbb{E}[X^2] + b\mathbb{E}[X] = \mathbb{E}[XY] \\ a\mathbb{E}[X] + b = \mathbb{E}[Y] \end{cases}$$

Sumando la primera ecuación y la segunda multiplicada por $-\mathbb{E}[X]$, se obtiene

$$a(\mathbb{E}[X^2] - \mathbb{E}[X]^2) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \iff a = \frac{\text{Cov}(X, Y)}{\mathbb{V}(X)}.$$

Sustituyendo el valor de a en la segunda y despejando b se obtiene

$$b = \mathbb{E}[Y] - \frac{\text{Cov}(X, Y)}{\mathbb{V}(X)} \mathbb{E}[X].$$

Por lo tanto, *la recta de regresión de Y basada en X* es

$$\begin{aligned}\hat{Y} &= \frac{\text{Cov}(X, Y)}{\mathbb{V}(X)} X + \mathbb{E}[Y] - \frac{\text{Cov}(X, Y)}{\mathbb{V}(X)} \mathbb{E}[X] \\ &= \frac{\text{Cov}(X, Y)}{\mathbb{V}(X)} (X - \mathbb{E}[X]) + \mathbb{E}[Y].\end{aligned}\tag{34}$$

Además el *error cuadrático medio* es igual a

$$\mathbb{E}[(Y - \hat{Y})^2] = \mathbb{V}(Y) (1 - \rho(X, Y)^2),\tag{35}$$

donde

$$\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}\tag{36}$$

es el llamado *coeficiente de correlación* de las variables X, Y .

Coeficiente de correlación

El coeficiente de correlación definido en (36) es la covarianza de las variables normalizadas

$$X^* := \frac{X - \mathbb{E}[X]}{\sigma(X)}, \quad Y^* := \frac{Y - \mathbb{E}[Y]}{\sigma(Y)}.\tag{37}$$

Este coeficiente es independiente de los orígenes y unidades de medida, esto es, para constantes a_1, a_2, b_1, b_2 con $a_1 > 0, a_2 > 0$, tenemos $\rho(a_1X + b_1, a_2Y + b_2) = \rho(X, Y)$.

Desafortunadamente, el término correlación sugiere implicaciones que no le son inherentes. Si X e Y son independientes, $\rho(X, Y) = 0$. Sin embargo la recíproca no es cierta. De hecho, *el coeficiente de correlación $\rho(X, Y)$ puede anularse incluso cuando Y es función de X* .

Ejemplo 3.2.

- Sea X una variable aleatoria que toma valores $\pm 1, \pm 2$ cada uno con probabilidad $\frac{1}{4}$ y sea $Y = X^2$. La distribución conjunta está dada por

$$p(-1, 1) = p(1, 1) = p(-2, 4) = p(2, 4) = 1/4.$$

Por razones de simetría ($\mathbb{E}[X] = 0$ y $\mathbb{E}[XY] = 0$) $\rho(X, Y) = 0$ incluso cuando Y es una función de X .

- Sean U y V variables *independientes* con la misma distribución, y sean $X = U + V$, $Y = U - V$. Entonces $\mathbb{E}[XY] = \mathbb{E}[U^2] - \mathbb{E}[V^2] = 0$ y $\mathbb{E}[Y] = 0$. En consecuencia, $\text{Cov}(X, Y) = 0$ y por lo tanto también $\rho(X, Y) = 0$. Por ejemplo, X e Y podrían ser la suma y la diferencia de los puntos de dos dados. Entonces X e Y son ambos pares ó ambos impares y por lo tanto dependientes.

Nota Bene. El coeficiente de correlación no es una medida general de la dependencia entre X e Y . Sin embargo, $\rho(X, Y)$ está conectado con la dependencia *lineal* de X e Y . En efecto, de la identidad (35) se deduce que $|\rho(X, Y)| \leq 1$ y que $\rho(X, Y) = \pm 1$ si y solo si Y es una función lineal de X (casí seguramente).

4. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Billingsley, P.: Probability and measure. John Wiley & Sons, New York. (1986)
2. Bertsekas, D. P., Tsitsiklis, J. N.: Introduction to Probability. M.I.T. Lecture Notes. (2000)
3. Durrett R.: Probability Theory and Examples. Duxbury Press, Belmont. (1996)
4. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons, New York. (1957)
5. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 2. John Wiley & Sons, New York. (1971)
6. Maronna R.: Probabilidad y Estadística Elementales para Estudiantes de Ciencias. Editorial Exacta, La Plata. (1995)
7. Ross, S.: Introduction to Probability Models. Academic Press, San Diego. (2007)

Ensayos Bernoulli y otras cositas
(Borradores, Curso 23)

Sebastian Grynberg

15-17 de abril de 2013



Jakob Bernoulli (1654 - 1705)

*En la “buena” te encontré
y en la “mala” te perdí ...
(Enrique Cadícamo)*

Índice

1. Ensayos Bernoulli	3
1.1. La distribución binomial: cantidad de éxitos en n ensayos	4
1.2. Término central	6
1.3. La distribución geométrica: tiempo de espera hasta el primer éxito	6
1.4. La distribución Pascal: tiempo de espera hasta el k -ésimo éxito	8
1.5. La distribución multinomial	9
1.6. Miselánea de ejemplos	10
2. La distribución de Poisson	12
2.1. Motivación: Aproximación de Poisson de la distribución binomial	12
2.2. La distribución Poisson	14
2.3. La aproximación Poisson. (Técnica de acoplamiento)	16
3. Cuentas con exponenciales	20
3.1. Motivación: pasaje de lo discreto a lo continuo	20
3.2. Distribución exponencial	21
3.3. Suma de exponenciales independientes de igual intensidad	21
3.4. Mínimos	22
4. Bibliografía consultada	24

1. Ensayos Bernoulli

Se trata de ensayos repetidos en forma independiente en los que hay sólo dos resultados posibles, usualmente denominados “éxito” y “fracaso”, cuyas probabilidades, p y $1 - p$, se mantienen constantes a lo largo de todos los ensayos.

El espacio muestral de cada ensayo individual está formado por dos puntos S y F . El espacio muestral de n ensayos Bernoulli contiene 2^n puntos o secuencias de n símbolos S y F , cada punto representa un resultado posible del experimento compuesto. Como los ensayos son independientes las probabilidades se multiplican. En otras palabras, *la probabilidad de cada sucesión particular es el producto que se obtiene reemplazando los símbolos S y F por p y 1 - p, respectivamente. Así,*

$$\mathbb{P}(SSFSF \dots FFS) = pp(1-p)p(1-p) \cdots (1-p)(1-p)p.$$

Ejemplo 1.1. Si repetimos en forma independiente un experimento aleatorio y estamos interesados en la ocurrencia del evento A al que consideramos “éxito”, tenemos ensayos Bernoulli con $p = \mathbb{P}(A)$. \square

Modelando ensayos Bernoulli. Los ensayos Bernoulli (con probabilidad de éxito p) se describen mediante una sucesión de variables aleatorias independientes e idénticamente distribuidas ($X_i : i \in \mathbb{N}$) cada una con distribución Bernoulli(p),

$$\mathbb{P}(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}, \quad x_i \in \{0, 1\}. \quad (1)$$

Esto es, $\mathbb{P}(X_i = 1) = p$ y $\mathbb{P}(X_i = 0) = 1 - p$. En este contexto, $X_i = 1$ significa que “el resultado del i -ésimo ensayo es éxito”.

Preguntas elementales. Se pueden formular varios tipos de preguntas relacionadas con los ensayos Bernoulli. Las más sencillas son las siguientes:

- (a) ¿Cuál es la cantidad total de éxitos en los primeros n ensayos?
- (b) ¿En n ensayos, cuál es el número de éxitos más probable?
- (c) ¿Cuánto “tiempo” hay que esperar para observar el primer éxito?
- (d) ¿Cuánto “tiempo” hay que esperar para observar el k -ésimo éxito?

En lo que sigue expresaremos las preguntas (a)-(d) en términos de las variables aleatorias X_i , $i \geq 1$, que describen los ensayos Bernoulli.

La *cantidad de éxitos en los primeros n ensayos* se describe mediante la suma de las primeras variables X_1, \dots, X_n

$$S_n := \sum_{i=1}^n X_i. \quad (2)$$

La pregunta (a) interroga por la distribución de probabilidades de la variable aleatoria S_n definida en (2). Esto es, para cada $k = 0, \dots, n$, se trata de determinar cuánto valen las probabilidades $\mathbb{P}(S_n = k)$. En cambio, la pregunta (b) interroga por el valor de k que maximiza a la función de k , $\mathbb{P}(S_n = k)$.

El *tiempo de espera hasta el primer éxito* se describe mediante la variable aleatoria

$$T_1 := \min\{i \in \mathbb{N} : X_i = 1\}, \quad (3)$$

y en general, el *tiempo de espera hasta el k -ésimo éxito*, $k \geq 1$ se describe, recursivamente, mediante

$$T_k := \min\{i > T_{k-1} : X_i = 1\}. \quad (4)$$

La pregunta (c) interroga por la distribución de probabilidades de la variable T_1 definida en (3): cuánto valen las probabilidades $\mathbb{P}(T_1 = n)$, $n \in \mathbb{N}$? Finalmente, la pregunta (d) interroga por la distribución de probabilidades de las variables T_k , $k \geq 2$, definidas en (4): cuánto valen las probabilidades $\mathbb{P}(T_k = n)$, $n \geq k$?

1.1. La distribución binomial: cantidad de éxitos en n ensayos

La cantidad de éxitos puede ser $0, 1, \dots, n$. El primer problema es determinar las correspondientes probabilidades. El evento *en n ensayos resultaron k éxitos y $n - k$ fracasos*

$$\left\{ (X_1, \dots, X_n) = (x_1, \dots, x_n) : \sum_{i=1}^n x_i = k \right\}$$

puede ocurrir de tantas formas distintas como k símbolos 1 se puedan ubicar en n lugares. En otras palabras, el evento considerado contiene $\binom{n}{k}$ puntos, cada uno de probabilidad

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=1}^n \{X_i = x_i\}\right) &= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} \\ &= p^k (1-p)^{n-k}. \end{aligned}$$

Por lo tanto,

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad 0 \leq k \leq n. \quad (5)$$

En particular, la probabilidad de que no ocurra ningún éxito en n ensayos es $(1-p)^n$ y la probabilidad de que ocurra al menos un éxito es $1 - (1-p)^n$.

La distribución de S_n , determinada en (5), se denomina *la distribución binomial de parámetros n y p* y se denota $\text{Binomial}(n, p)$.

Nota Bene. Por definición, la distribución binomial de parámetros n y p es *la distribución de una suma de n variables aleatorias independientes cada con distribución Bernoulli de parámetro p* .

Ejemplo 1.2. Se tira un dado equilibrado 11 veces y en cada tiro se apuesta al 6, ¿cuál es la probabilidad de ganar exactamente 2 veces? Como el dado es equilibrado, la probabilidad de éxito es $1/6$ y la cantidad de éxitos en 11 tiros tiene distribución Binomial $(11, 1/6)$. Por lo tanto, la probabilidad requerida es

$$\binom{11}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^9 = 0.2960\dots$$

□

Ejemplo 1.3. Cada artículo producido por una máquina será defectuoso con probabilidad 0.1, independientemente de los demás. En una muestra de 3, ¿cuál es la probabilidad de encontrar a lo sumo un defectuoso?

Si X es la cantidad de artículos defectuosos en la muestra, entonces $X \sim \text{Binomial}(3, 0.1)$. En consecuencia,

$$\mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = \binom{3}{0} (0.1)^0 (0.9)^3 + \binom{3}{1} (0.1)^1 (0.9)^2 = 0.972.$$

□

Ejemplo 1.4. Un avión se mantendrá en vuelo mientras funcionen al menos el 50% de sus motores. Si cada motor del avión en vuelo puede fallar con probabilidad $1 - p$ independientemente de los demás, ¿para cuáles valores de $p \in (0, 1)$ es más seguro un avión de 4 motores que uno de 2?

Como cada motor puede fallar o funcionar independientemente de los demás, la cantidad de motores que siguen funcionando es una variable aleatoria con distribución binomial. La probabilidad de que un avión de 4 motores realice un vuelo exitoso es

$$\binom{4}{2} p^2 (1-p)^2 + \binom{4}{3} p^3 (1-p) + \binom{4}{4} p^4 = 6p^2(1-p)^2 + 4p^3(1-p) + p^4,$$

mientras que la correspondiente probabilidad para un avión de 2 motores es

$$\binom{2}{1} p(1-p) + \binom{2}{2} p^2 = 2p(1-p) + p^2.$$

En consecuencia, el avión de 4 motores es más seguro que el de 2 si

$$6p^2(1-p)^2 + 4p^3(1-p) + p^4 > 2p(1-p) + p^2$$

lo que es equivalente a las siguientes expresiones simplificadas

$$3p^3 - 8p^2 + 7p - 2 > 0 \iff 3(p - 2/3)(p - 1)^2 > 0 \iff p > 2/3.$$

Por lo tanto, el avión de 4 motores es más seguro cuando la probabilidad de que cada motor se mantenga en funcionamiento es mayor que $2/3$, mientras que el avión de 2 motores es más seguro cuando esa probabilidad es menor que $2/3$. □

Ejemplo 1.5. Si la probabilidad de éxito es $p = 0.01$, cuántos ensayos se deben realizar para asegurar que la probabilidad de que ocurra por lo menos un éxito sea al menos $1/2$?

Buscamos el menor entero n tal que $1 - (0.99)^n \geq \frac{1}{2}$, o equivalentemente $\frac{1}{2} \geq (0.99)^n$. Tomando logaritmos $-\log 2 \geq n \log(0.99)$ y despejando n resulta $n \geq -\log(2)/\log(0.99) \approx 68.96$. Por lo tanto, $n = 69$. □

1.2. Término central

De la fórmula (5) se puede ver que

$$\begin{aligned} \frac{\mathbb{P}(S_n = k)}{\mathbb{P}(S_n = k-1)} &= \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\binom{n}{k-1} p^{k-1} (1-p)^{n-k+1}} = \frac{(k-1)!(n-k+1)!p}{k!(n-k)!(1-p)} \\ &= \frac{(n-k+1)p}{k(1-p)} = 1 + \frac{(n+1)p - k}{k(1-p)}. \end{aligned} \quad (6)$$

De (6) se deduce que $\mathbb{P}(S_n = k)$ crece cuando $k < (n+1)p$ y decrece cuando $k > (n+1)p$. Si $(n+1)p$ es un número entero, entonces $\mathbb{P}(S_n = (n+1)p) = \mathbb{P}(S_n = (n+1)p-1)$. En otras palabras, la cantidad más probable de éxitos en n ensayos es $m := [(n+1)p]$. Salvo en el caso en que $m = (n+1)p$, donde también lo es $m-1$.

Cuando $p = \frac{1}{2}$ el resultado anterior se puede observar directamente en el triángulo de Pascal: en el centro de las filas pares está el máximo. En la región central de las filas impares hay dos máximos.

Ejemplo 1.6. Se tira un dado equilibrado n veces y en cada tiro se apuesta al 6. ¿Cuál es la cantidad más probable de éxitos cuando $n = 12$? y cuando $n = 11$?

La cantidad de éxitos tiene distribución Binomial (n, p) , donde $p = 1/6$. Cuando $n = 12$, $(n+1)p = 13/6 = 2.16\dots$ y entonces la cantidad más probable de éxitos es $m = 2$. Cuando $n = 11$, $(n+1)p = 2$ y entonces la cantidad más probable de éxitos es $m = 1$ o $m = 2$. \square

1.3. La distribución geométrica: tiempo de espera hasta el primer éxito

El tiempo que hay que esperar para observar el primer éxito en una sucesión de ensayos Bernoulli puede ser $n = 1, 2, \dots$. El evento $T_1 = 1$ significa que se obtuvo éxito en el primer ensayo y tiene probabilidad p . Para cada $n \geq 2$, el evento $T_1 = n$ significa que en los primeros $n-1$ ensayos se obtuvieron fracasos y que en el n -ésimo se obtuvo éxito, lo que tiene probabilidad $(1-p)^{n-1}p$. Por lo tanto, la distribución de T_1 es

$$\mathbb{P}(T_1 = n) = (1-p)^{n-1}p, \quad n \in \mathbb{N}. \quad (7)$$

El evento $T_1 > n$ significa que los primeros n ensayos de la sucesión resultaron fracaso. Por lo tanto,

$$\mathbb{P}(T_1 > n) = (1-p)^n, \quad n \geq 1. \quad (8)$$

La distribución de T_1 se denomina *distribución geométrica de parámetro p* y se designa mediante Geométrica(p).

Ejemplo 1.7. Se arroja repetidamente un dado equilibrado. ¿Cuál es la probabilidad de que el primer 6 aparezca antes del quinto tiro?. La probabilidad de obtener 6 es $1/6$ y la cantidad de tiros hasta obtener el primer 6 tiene distribución Geométrica($1/6$). Por lo tanto, la probabilidad requerida es

$$1/6 + (5/6)(1/6) + (5/6)^2(1/6) + (5/6)^3(1/6) = (1/6) \left(\frac{1 - (5/6)^4}{1 - (5/6)} \right) = 1 - (5/6)^4 = 0.5177\dots$$

\square

Ejemplo 1.8 (Ocurrencias casi seguras). Si al realizarse un experimento aleatorio un evento A tiene probabilidad positiva de ocurrir, entonces en una sucesión de experimentos independientes el evento A ocurrirá *casi seguramente*.

En efecto, el tiempo de espera hasta que ocurra el evento A es una variable aleatoria T_A con distribución geométrica de parámetro $p = \mathbb{P}(A)$. Si se observa que

$$\{T_A > 1\} \supseteq \{T_A > 2\} \supseteq \{T_A > 3\} \supseteq \dots$$

y que

$$\{T_A = \infty\} = \bigcap_{n \geq 1} \{T_A > n\}$$

y se usa la propiedad de continuidad de \mathbb{P} , se obtiene que

$$\mathbb{P}(T_A = \infty) = P\left(\bigcap_{n \geq 1} \{T_A > n\}\right) = \lim_{n \rightarrow \infty} \mathbb{P}(T_A > n) = \lim_{n \rightarrow \infty} (1-p)^n = 0.$$

Por lo tanto, $\mathbb{P}(T_A < \infty) = 1$. □

Pérdida de memoria

La variable aleatoria, T , con distribución geométrica de parámetro p tiene la propiedad de *pérdida de memoria*,

$$\mathbb{P}(T > n+m | T > n) = \mathbb{P}(T > m) \quad n, m \in \mathbb{N} \quad (9)$$

La identidad (9) se obtiene de (8) y de la fórmula de probabilidad condicional:

$$\begin{aligned} \mathbb{P}(T > n+m | T > n) &= \frac{\mathbb{P}(T > n+m, T > n)}{\mathbb{P}(T > n)} \\ &= \frac{\mathbb{P}(T > n+m)}{\mathbb{P}(T > n)} = \frac{(1-p)^{n+m}}{(1-p)^n} \\ &= (1-p)^m = \mathbb{P}(T > m). \end{aligned}$$

De hecho, la propiedad de pérdida de memoria definida en (9) caracteriza a la distribución geométrica.

Teorema 1.9. Si T es una variable aleatoria a valores en \mathbb{N} con la propiedad de pérdida de memoria, entonces $T \sim \text{Geométrica}(p)$, donde $p = \mathbb{P}(T = 1)$.

Demostración. Sea $G(n) := \mathbb{P}(T > n)$. Si T pierde memoria, tenemos que

$$G(n+m) = G(n)G(m) \quad (10)$$

De (10) sigue que $G(2) = G(1)G(1) = G(1)^2$, $G(3) = G(2)G(1) = G(1)^3$ y en general $G(n) = G(1)^n$ cualquiera sea $n \in \mathbb{N}$. En otros términos, la distribución de T es tal que

$$\mathbb{P}(T > n) = G(1)^n.$$

Por lo tanto,

$$\mathbb{P}(T = n) = \mathbb{P}(T > n-1) - \mathbb{P}(T > n) = G(1)^{n-1} - G(1)^n = G(1)^{n-1}(1 - G(1)).$$

□

1.4. La distribución Pascal: tiempo de espera hasta el k -ésimo éxito

Si se quieren observar k -éxitos en una sucesión de ensayos Bernoulli lo mínimo que se debe esperar es k ensayos. ¿Cuándo ocurre el evento $T_k = n$, $n \geq k$? El n -ésimo ensayo debe ser éxito y en los $n - 1$ ensayos anteriores deben ocurrir exactamente $k - 1$ éxitos. Hay $\binom{n-1}{k-1}$ formas distintas de ubicar $k - 1$ símbolos 1 en $n - 1$ lugares. Por lo tanto,

$$\mathbb{P}(T_k = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \quad n \geq k. \quad (11)$$

La distribución de T_k se denomina *distribución Pascal de parámetros k y p* y se designa mediante $\text{Pascal}(k, p)$.

La distribución Pascal de parámetros k y p es la distribución de una suma de k variables aleatorias independientes cada una con ley Geométrica(p). Lo cual es intuitivamente claro si se piensa en el modo que arribamos a su definición.

En efecto, definiendo $T_0 := 0$ vale que

$$T_k = \sum_{i=1}^k (T_i - T_{i-1}).$$

Basta ver que para cada $i = 1, \dots, k$ las diferencias $T_i - T_{i-1}$ son independientes y todas se distribuyen como $T_1 \sim \text{Geométrica}(p)$. De acuerdo con la regla del producto

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=1}^k \{T_i - T_{i-1} = m_i\}\right) &= \mathbb{P}(T_1 = m_1) \\ &\times \prod_{i=2}^{n-1} \mathbb{P}\left(T_i - T_{i-1} = m_i \mid \bigcap_{j=1}^{i-1} \{T_j - T_{j-1} = m_j\}\right). \end{aligned} \quad (12)$$

Si se sabe que $T_1 = m_1, \dots, T_{i-1} - T_{i-2} = m_{i-1}$, entonces el evento $T_i - T_{i-1} = m_i$ depende las variables aleatorias $X_{\sum_{j=1}^{i-1} m_j + 1}, \dots, X_{\sum_{j=1}^i m_j}$ y equivale a decir que las primeras $m_i - 1$ de esas variables valen 0 y la última vale 1. En consecuencia,

$$\mathbb{P}\left(T_i - T_{i-1} = m_i \mid \bigcap_{j=1}^{i-1} \{T_j - T_{j-1} = m_j\}\right) = (1-p)^{m_i-1} p. \quad (13)$$

De (12) y (13) se deduce que

$$\mathbb{P}\left(\bigcap_{i=1}^k \{T_i - T_{i-1} = m_i\}\right) = \prod_{i=1}^k (1-p)^{m_i-1} p. \quad (14)$$

De la factorización (14) se deduce que $T_1, T_2 - T_1, \dots, T_k - T_{k-1}$ son independientes y que cada una tiene distribución geométrica de parámetro p . \square

Ejemplo 1.10. Lucas y Monk disputan la final de un campeonato de ajedrez. El primero que gane 6 partidas (no hay tablas) resulta ganador. La probabilidad de que Lucas gane cada partida es $3/4$. ¿Cuál es la probabilidad de que Lucas gane el campeonato en la novena partida? La cantidad de partidas que deben jugarse hasta que Lucas gane el campeonato tiene distribución $\text{Pascal}(6, 3/4)$. Por lo tanto, la probabilidad requerida es

$$\binom{8}{5} \left(\frac{3}{4}\right)^6 \left(\frac{1}{4}\right)^3 = 0.1557\dots$$

\square

Ejemplo 1.11. En una calle hay tres parquímetros desocupados. Se estima que en los próximos 10 minutos pasarán 6 coches por esa calle y, en media, el 80% tendrá que estacionarse en alguno de ellos. Calcular la probabilidad de que los tres parquímetros sean ocupados en los próximos 10 minutos.

La probabilidad requerida es la probabilidad de que la cantidad, N , de ensayos hasta el tercer éxito sea menor o igual que 6. Como N tiene distribución $\text{Pascal}(3, 0.8)$ resulta que

$$\begin{aligned}\mathbb{P}(N \leq 6) &= \sum_{n=3}^6 \mathbb{P}(N = n) = \sum_{n=3}^6 \binom{n-1}{2} (0.8)^3 (0.2)^{n-3} \\ &= (0.8)^3 \left[\binom{2}{2} (0.2)^0 + \binom{3}{2} (0.2)^1 + \binom{4}{2} (0.2)^2 + \binom{5}{2} (0.2)^3 \right] \\ &= (0.8)^3 [1 + 3(0.2) + 6(0.2)^2 + 10(0.2)^3] \\ &= 0.983\dots\end{aligned}$$

Notar que una forma alternativa de obtener el mismo resultado es sumar las probabilidades de observar 3, 4, 5, 6 éxitos en 6 ensayos Bernoulli. \square

Relación entre las distribuciones Binomial y Pascal. Sean $S_n \sim \text{Binomial}(n, p)$ y $T_k \sim \text{Pascal}(k, p)$. Vale que

$$\mathbb{P}(S_n \geq k) = \mathbb{P}(T_k \leq n). \quad (15)$$

En efecto, decir que en n ensayos Bernoulli ocurren por lo menos k éxitos es lo mismo que decir que el tiempo de espera hasta observar el k -ésimo éxito no supera a n . \square

1.5. La distribución multinomial

La distribución binomial se puede generalizar al caso de n ensayos independientes donde cada ensayo puede tomar uno de varios resultados. Sean $1, 2, \dots, r$ los resultados posibles de cada ensayo y supongamos que para cada $k \in \{1, 2, \dots, r\}$ la probabilidad p_k de observar el valor k se mantiene constante a lo largo de los ensayos. La pregunta es: ¿Cuántas veces ocurre cada uno de los resultados en los primeros n ensayos?

Consideramos una sucesión X_1, X_2, \dots de variables aleatorias independientes e idénticamente distribuidas a valores $\{1, 2, \dots, r\}$ tal que $\mathbb{P}(X_i = k) = p_k$. Fijado n , para cada $k = 1, \dots, r$ definimos la variables $M_k = \sum_{i=1}^n \mathbf{1}\{X_i = k\}$. La variable M_k cuenta la cantidad de veces que ocurre el resultado k en n ensayos. La probabilidad de que en n ensayos el resultado 1 ocurra m_1 veces, el resultado 2 ocurra m_2 veces, etc. es

$$\mathbb{P}(M_1 = m_1, M_2 = m_2, \dots, M_r = m_r) = \frac{n!}{m_1! m_2! \cdots m_r!} p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r}, \quad (16)$$

donde los m_k son enteros no negativos sujetos a la condición $m_1 + m_2 + \cdots + m_r = n$.

Si $r = 2$, entonces (16) se reduce a la distribución Binomial con $p_1 = p$, $p_2 = 1 - p$, $k_1 = k$ y $k_2 = n - k$.

1.6. ♣ Miscelánea de ejemplos

Observación 1.12 (Desarrollo de Taylor). Para todo $x \in (0, 1)$ vale que

$$\frac{1}{(1-x)^{k+1}} = \sum_{n \geq 0} \binom{n+k}{k} x^n. \quad (17)$$

La identidad (17) se obtiene desarrollando la función $h(x) = (1-x)^{-(k+1)}$ en serie de Taylor alrededor del 0: observando que $h^{(n)}(0) = (k+1)(k+2)\cdots(k+n)$, se obtiene que $\frac{h^{(n)}(0)}{n!} = \binom{n+k}{k}$. \square

Ejemplo 1.13 (Variable compuesta). Sean $N_1; X_1, X_2, \dots$ una sucesión de variables aleatorias independientes. Supongamos que $N_1 \sim \text{Geométrica}(p_1)$ y que $X_i \sim \text{Bernoulli}(p_2)$, $i \geq 1$. Entonces,

$$N_2 = \sum_{i=1}^{N_1-1} X_i \sim \text{Geométrica} \left(\frac{p_1}{p_1 + p_2(1-p_1)} \right) - 1. \quad (18)$$

Por definición $N_2|N_1 = n \sim \text{Binomial}(n-1, p_2)$. Aplicando la fórmula de probabilidad total obtenemos

$$\begin{aligned} \mathbb{P}(N_2 = k) &= \sum_{n \geq 1} \mathbb{P}(N_2 = k | N_1 = n) \mathbb{P}(N_1 = n) \\ &= \sum_{n \geq k+1} \binom{n-1}{k} p_2^k (1-p_2)^{n-1-k} (1-p_1)^{n-1} p_1 \\ &= \sum_{m \geq 0} \binom{m+k}{k} p_2^k (1-p_2)^m (1-p_1)^{m+k} p_1 \\ &= (p_2(1-p_1))^k p_1 \sum_{m \geq 0} \binom{n+k}{k} [(1-p_1)(1-p_2)]^m. \end{aligned} \quad (19)$$

Usando (17) vemos que

$$\begin{aligned} \sum_{m \geq 0} \binom{m+k}{k} [(1-p_1)(1-p_2)]^m &= \frac{1}{(1-(1-p_1)(1-p_2))^{k+1}} \\ &= \frac{1}{(p_1 + p_2(1-p_1))^{k+1}}. \end{aligned} \quad (20)$$

Combinando (19) y (20) obtenemos que

$$\mathbb{P}(N_2 = k) = \frac{(p_2(1-p_1))^k p_1}{(p_1 + p_2(1-p_1))^{k+1}} = \left(\frac{p_2(1-p_1)}{p_1 + p_2(1-p_1)} \right)^k \left(\frac{p_1}{p_1 + p_2(1-p_1)} \right). \quad (21)$$

\square

Ejemplo 1.14 (Rachas). Para cada número entero $m > 1$ sea Y_m la cantidad de ensayos Bernoulli(p) que se deben realizar hasta obtener por primera vez una racha de m éxitos seguidos. En lo que sigue vamos a calcular $E[Y_m]$ mediante condicionales. Para ello introducimos

una variable aleatoria auxiliar N que cuenta la cantidad de ensayos que deben realizarse hasta obtener por primera vez un *fracaso* y usaremos la identidad $\mathbb{E}[Y_m] = \mathbb{E}[\mathbb{E}[Y_m|N]]$.

Observando que

$$Y_m|N = n \sim \begin{cases} n + Y_m & \text{si } n \leq m, \\ m & \text{si } n > m, \end{cases}$$

obtenemos la expresión de la función de regresión

$$\varphi(n) = \mathbb{E}[Y_m|N = n] = \begin{cases} n + \mathbb{E}[Y_m] & \text{si } n \leq m, \\ m & \text{si } n > m. \end{cases}$$

En consecuencia, $\mathbb{E}[Y_m|N] = N\mathbf{1}\{N \leq m\} + \mathbb{E}[Y_m]\mathbf{1}\{N \leq m\} + m\mathbf{1}\{N > m\}$, de donde se deduce que $\mathbb{E}[Y_m] = \mathbb{E}[N\mathbf{1}\{N \leq m\}] + \mathbb{E}[Y_m]\mathbb{P}(N \leq m) + m\mathbb{P}(N > m)$. Equivalentemente,

$$\mathbb{E}[Y_m] = \frac{\mathbb{E}[N\mathbf{1}\{N \leq m\}]}{\mathbb{P}(N > m)} + m. \quad (22)$$

Debido a que $N\mathbf{1}\{N \leq m\} = N - N\mathbf{1}\{N > m\}$ el primer término del lado derecho de la igualdad (22) se puede expresar de siguiente forma

$$\begin{aligned} \frac{\mathbb{E}[N\mathbf{1}\{N \leq m\}]}{\mathbb{P}(N > m)} &= \frac{\mathbb{E}[N] - \mathbb{E}[N\mathbf{1}\{N > m\}]}{\mathbb{P}(N > m)} = \frac{\mathbb{E}[N]}{\mathbb{P}(N > m)} - \mathbb{E}[N|\mathbb{P}(N > m)] \\ &= \frac{\mathbb{E}[N]}{\mathbb{P}(N > m)} - \mathbb{E}[N] - m. \end{aligned} \quad (23)$$

La última igualdad se deduce de la propiedad de pérdida de memoria de la distribución Geométrica. De $N|\mathbb{P}(N > m) \sim m + N$, resulta que $\mathbb{E}[N|\mathbb{P}(N > m)] = m + \mathbb{E}[N]$.

Combinando (22) y (23) obtenemos

$$\mathbb{E}[Y_m] = \frac{\mathbb{E}[N]}{\mathbb{P}(N > m)} - \mathbb{E}[N] = \frac{\mathbb{E}[N]\mathbb{P}(N \leq m)}{\mathbb{P}(N > m)} = \frac{1 - p^m}{(1 - p)p^m}. \quad (24)$$

□

Ejemplo 1.15 (Coleccionista I). Sea M una variable aleatoria a valores $1, 2, \dots, m$. Sea $(M_n : n \in \mathbb{N})$ una sucesión de variables aleatorias independientes tal que $M_n \sim M$ para todo $n \in \mathbb{N}$. Sea $K = \min\{n \geq m : \{M_1, \dots, M_n\} = \{1, 2, \dots, m\}\}$ el tamaño de muestra mínimo que se necesita para “coleccionar” todos los valores $1, 2, \dots, m$. En lo que sigue vamos a calcular $\mathbb{E}[K]$ mediante condicionales. Introducimos un elemento aleatorio C que indica el orden en que se obtuvieron los valores $1, 2, \dots, m$ y usamos la identidad $\mathbb{E}[K] = \mathbb{E}[\mathbb{E}[K|C]]$.

Sea $S(m)$ al conjunto de todas las permutaciones de los números $1, 2, \dots, m$. Para cada permutación $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_m) \in S(m)$ vale que:

$$\mathbb{P}(C = \sigma) = \prod_{k=1}^{m-1} \frac{\mathbb{P}(M = \sigma_k)}{\sum_{i=k}^m \mathbb{P}(M = \sigma_i)}.$$

Por otra parte

$$K|C = \sigma \sim 1 + \sum_{k=1}^{m-1} N(\sigma_i : 1 \leq i \leq k),$$

donde $N(\sigma_i : 1 \leq i \leq k) \sim \text{Geométrica}(\sum_{i=k+1}^m \mathbb{P}(M = \sigma_i))$. Por lo tanto,

$$\begin{aligned}\mathbb{E}[K] &= \sum_{\sigma \in S(m)} \mathbb{E}[K|C = \sigma] \mathbb{P}(C = \sigma) \\ &= \sum_{\sigma \in S(m)} \left(1 + \sum_{k=1}^{m-1} \frac{1}{\sum_{i=k+1}^m \mathbb{P}(M = \sigma_i)} \right) \prod_{k=1}^{m-1} \frac{\mathbb{P}(M = \sigma_k)}{\sum_{i=k}^m \mathbb{P}(M = \sigma_i)}.\end{aligned}\quad (25)$$

En el caso particular en que $\mathbb{P}(M = i) = 1/m$ para todo $i \in \{1, 2, \dots, m\}$ tenemos que

$$\begin{aligned}\mathbb{E}[K] &= \sum_{\sigma \in S(m)} \left(1 + \sum_{k=1}^{m-1} \frac{1}{\sum_{i=k+1}^m 1/m} \right) \prod_{k=1}^{m-1} \frac{1/m}{\sum_{i=k}^m 1/m} \\ &= m! \left(1 + \sum_{k=1}^{m-1} \frac{1}{\sum_{i=k+1}^m 1/m} \right) \frac{1}{m!} = \sum_{k=0}^{m-1} \frac{1}{\sum_{i=k+1}^m 1/m} = m \sum_{i=1}^m \frac{1}{i}.\end{aligned}\quad (26)$$

□

Ejemplo 1.16 (Coleccionista II). Sea X_1, X_2, \dots una sucesión de variables aleatorias independientes e idénticamente distribuidas a valores $1, 2, \dots, r$. Sea $N_r = \min\{n \geq 1 : X_n = r\}$. Para cada $i = 1, \dots, r-1$ sea $M_i = \sum_{n=1}^{N_r-1} \mathbf{1}\{X_n = i\}$. Queremos hallar la función de probabilidad de M_i .

Por definición $N_r \sim \text{Geométrica}(p_r)$ y $M_i|N_r = n \sim \text{Binomial}(n-1, p_i(1-p_r)^{-1})$. De acuerdo con el Ejemplo 1.13 tenemos que

$$M_i \sim \text{Geométrica} \left(\frac{p_r}{p_r + p_i(1-p_r)^{-1}(1-p_r)} \right) - 1 = \text{Geométrica} \left(\frac{p_r}{p_r + p_i} \right) - 1.$$

En particular, $\mathbb{E}[M_i] = p_i/p_r$ y $\mathbb{V}(M_i) = p_i(p_r + p_i)/p_r^2$. □

2. La distribución de Poisson

2.1. Motivación: Aproximación de Poisson de la distribución binomial

En diversas aplicaciones tenemos que tratar con ensayos Bernoulli donde, para decirlo de algún modo, n es grande y p es pequeño, mientras que el producto $\lambda = np$ es moderado. En tales casos conviene usar una aproximación de las probabilidades $\mathbb{P}(S_n = k)$, donde $S_n \sim \text{Binomial}(n, p)$ y $p = \lambda/n$. Para $k = 0$ tenemos

$$\mathbb{P}(S_n = 0) = (1-p)^n = \left(1 - \frac{\lambda}{n} \right)^n. \quad (27)$$

Tomando logaritmos y usando el desarrollo de Taylor,

$$\log(1-t) = -t - \frac{1}{2}t^2 - \frac{1}{3}t^3 - \frac{1}{4}t^4 - \dots,$$

se obtiene

$$\log \mathbb{P}(S_n = 0) = n \log \left(1 - \frac{\lambda}{n} \right) = -\lambda - \frac{\lambda^2}{2n} - \dots \quad (28)$$

En consecuencia, para n grande se tiene que

$$\mathbb{P}(S_n = 0) \approx e^{-\lambda}, \quad (29)$$

donde el signo \approx se usa para indicar una igualdad aproximada (en este caso de orden de magnitud $1/n$). Más aún, usando la identidad (6) se puede ver que para cada k fijo y n suficientemente grande

$$\frac{\mathbb{P}(S_n = k)}{\mathbb{P}(S_n = k - 1)} = \frac{(n - k + 1)p}{k(1 - p)} \approx \frac{\lambda}{k}. \quad (30)$$

Recursivamente se concluye que

$$\begin{aligned} \mathbb{P}(S_n = 1) &\approx \lambda \cdot \mathbb{P}(S_n = 0) \approx \lambda e^{-\lambda}, \\ \mathbb{P}(S_n = 2) &\approx \frac{\lambda}{2} \cdot \mathbb{P}(S_n = 1) \approx \frac{\lambda^2}{2} e^{-\lambda}, \end{aligned}$$

y en general

$$\mathbb{P}(S_n = k) \approx \frac{\lambda^k}{k!} e^{-\lambda}. \quad (31)$$

La igualdad aproximada (31) se llama *la aproximación de Poisson de la distribución binomial*.

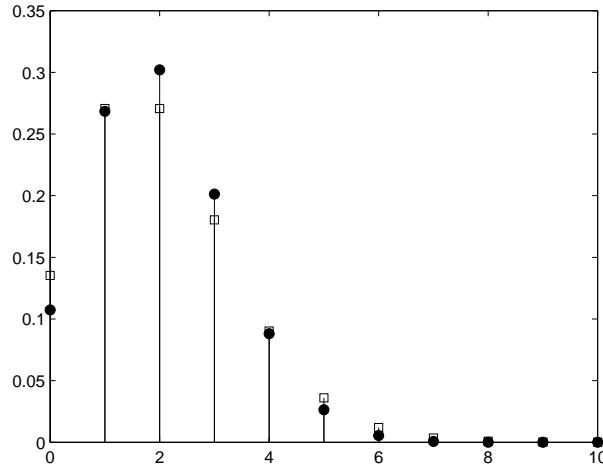


Figura 1: Comparación. Funciones de probabilidad de las distribuciones Binomial($10, 1/5$) (bolita negra) y Poisson(2) (cuadradillo vacío).

Otro modo de obtener el mismo resultado.

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{k!} \left(\frac{np}{1-p} \right)^k (1-p)^{n-k} \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}.$$

Ejemplo 2.1 (Artículos defectuosos). Una industria produce tornillos. Supongamos que la probabilidad de que un tornillo resulte defectuoso sea $p = 0.015$, entonces la probabilidad de que una caja de 100 tornillos no contenga ninguno defectuoso es $(0.985)^{100} = 0.2206\dots$. La aproximación de Poisson es $e^{-1.5} = 0.2231\dots$ y es suficientemente próxima para la mayoría de los propósitos prácticos. Si se pregunta: Cuántos tornillos debería contener la caja para que la probabilidad de encontrar al menos 100 tornillos sin defectos sea 0.8 o mejor? Si $100 + x$ es el número buscado, entonces x es un número pequeño. Para aplicar la aproximación de Poisson para $n = 100 + x$ ensayos debemos poner $\lambda = np$, pero np es aproximadamente $100p = 1.5$. Buscamos el menor entero x para el cual

$$e^{-1.5} \left(1 + \frac{1.5}{1} + \dots + \frac{(1.5)^x}{x!} \right) \geq 0.8 \quad (32)$$

Para $x = 1$ el valor del lado izquierdo de la inecuación (32) es aproximadamente 0.558, para $x = 2$ es aproximadamente 0.809. Por lo tanto, la aproximación de Poisson permite concluir que se necesitan 102 tornillos. En realidad la probabilidad de encontrar al menos 100 tornillos sin defectos en una caja de 102 es 0.8022.... \square

2.2. La distribución Poisson

Sea $\lambda > 0$. Una variable aleatoria N tiene distribución Poisson(λ) si sus posibles valores son los enteros no negativos y si

$$\mathbb{P}(N = n) = e^{-\lambda} \frac{\lambda^n}{n!}, \quad n = 0, 1, \dots \quad (33)$$

Media y varianza. Usando el desarrollo de Taylor de la función exponencial $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ se demuestra que $\mathbb{E}[N] = \lambda$ y $\mathbb{V}(N) = \lambda$.

Aditividad. El rasgo más importante de la distribución Poisson es su aditividad.

Teorema 2.2 (Aditividad). *Si N_1 y N_2 son variables aleatorias independientes con distribución Poisson de medias λ_1 y λ_2 , respectivamente. Entonces,*

$$N_1 + N_2 \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

Demostración.

$$\begin{aligned} \mathbb{P}(N_1 + N_2 = n) &= \sum_{m=0}^n \mathbb{P}(N_1 = m, N_2 = n - m) = \sum_{m=0}^n \mathbb{P}(N_1 = m) \mathbb{P}(N_2 = n - m) \\ &= \sum_{m=0}^n e^{-\lambda_1} \frac{\lambda_1^m}{m!} e^{-\lambda_2} \frac{\lambda_2^{n-m}}{(n-m)!} = \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{m=0}^n \binom{n}{m} \lambda_1^m \lambda_2^{n-m} \\ &= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^n}{n!}. \end{aligned}$$

\square

Nota Bene. El resultado del Teorema 2.2 se extiende por inducción a la suma de una cantidad finita de variables aleatorias independientes con distribución Poisson.

Teorema 2.3 (Competencia). *Sean N_1, N_2, \dots, N_m variables aleatorias independientes, cada N_j con distribución Poisson de media λ_j , respectivamente. Sea $S = N_1 + \dots + N_m$. Entonces, para cada $n \geq 1$ vale que*

$$(N_1, N_2, \dots, N_m) | S = n \sim \text{Multinomial} \left(n, \frac{\lambda_1}{\lambda}, \frac{\lambda_2}{\lambda}, \dots, \frac{\lambda_m}{\lambda} \right),$$

donde $\lambda = \sum_j \lambda_j$. En particular,

$$\mathbb{P}(N_j = 1 | S = 1) = \frac{\lambda_j}{\lambda}.$$

Demostración. La suma $S = N_1 + \dots + N_m$ tiene distribución Poisson de media $\lambda = \sum_j \lambda_j$; y entonces siempre que $n_1 + \dots + n_m = n$,

$$\begin{aligned} \mathbb{P}(N_1 = n_1, \dots, N_m = n_m | S = n) &= \frac{\mathbb{P}(N_1 = n_1, \dots, N_m = n_m)}{\mathbb{P}(S = n)} \\ &= \prod_j \left(e^{-\lambda_j} \frac{\lambda_j^{n_j}}{n_j!} \right) / \left(e^{-\lambda} \frac{\lambda^n}{n!} \right) \\ &= \frac{n!}{n_1! n_2! \cdots n_m!} \prod_j \left(\frac{\lambda_j}{\lambda} \right)^{n_j}. \end{aligned}$$

□

Nota Bene. En el caso particular $n = 2$, el resultado del Teorema 2.3 se reduce a que, si N_1 y N_2 son variables aleatorias independientes con distribución Poisson de medias λ_1 y λ_2 , respectivamente, entonces, dado que $N_1 + N_2 = n$, la distribución condicional de N_1 es Binomial(n, p), donde $p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$. □

Teorema 2.4 (Adelgazamiento). *Sea N una variable aleatoria Poisson de media λ . Sea M una variable aleatoria tal que*

$$M | N = n \sim \text{Binomial}(n, p).$$

Entonces, M y $N - M$ son variables aleatorias independientes con distribución Poisson de medias $p\lambda$ y $(1-p)\lambda$, respectivamente.

Demostración. Sean $m, k \geq 0$

$$\begin{aligned} \mathbb{P}(M = m, N - M = k) &= \mathbb{P}(M = m, N - M = k | N = m + k) \mathbb{P}(N = m + k) \\ &= \mathbb{P}(M = m | N = m + k) \mathbb{P}(N = m + k) \\ &= \left(\binom{m+k}{m} p^m (1-p)^k \right) e^{-\lambda} \frac{\lambda^{m+k}}{(m+k)!} \\ &= \left(e^{-p\lambda} \frac{(p\lambda)^m}{m!} \right) \left(e^{-(1-p)\lambda} \frac{((1-p)\lambda)^k}{k!} \right). \end{aligned}$$

□

Ejercicios adicionales

1. Sea N una variable aleatoria con distribución Poisson de media λ . Mostrar que

$$\mathbb{P}(N = n) = \frac{\lambda}{n} \mathbb{P}(N = n - 1), \quad n = 1, 2, \dots$$

Usar ese resultado para encontrar el valor de n para el cual $\mathbb{P}(N = n)$ es maximal.

2. Se lanza una moneda una cantidad aleatoria N de veces, donde N tiene distribución Poisson. Sean N_1 y N_2 la cantidad de total de caras y de cecas observadas, respectivamente. Mostrar que las variables aleatorias N_1 y N_2 son independientes y que tienen distribución Poisson.

3. Sea X_1, X_2, \dots una sucesión de variables aleatorias independientes, cada una con distribución Bernoulli(p). Para cada $n \geq 1$ se define $S_n := \sum_{i=1}^n X_i$. Por convención, $S_0 := 0$. Sea N una variable aleatoria con distribución Poisson(λ). Mostrar que $S_N \sim \text{Poisson}(p\lambda)$.
-

2.3. La aproximación Poisson. (Técnica de acoplamiento)

En lo que sigue mostraremos que cuando se consideran una gran cantidad de eventos independientes y cada uno de ellos tiene una probabilidad muy pequeña de ocurrir, la cantidad de tales eventos que realmente ocurre tiene una distribución “cercana” a la distribución Poisson.

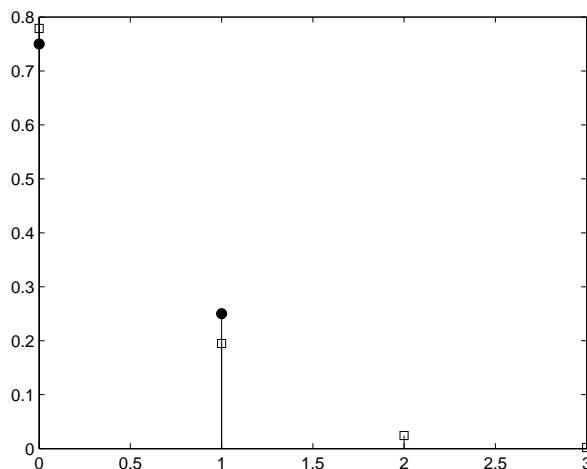


Figura 2: Comparación de las funciones de probabilidad de las distribuciones Bernoulli($1/4$) (bolita negra) y Poisson($1/4$) (cuadradillo vacío)

Construcción conjunta de variables Bernoulli y Poisson (Acoplamiento).

Para cada $p \in [0, 1]$ dividimos el intervalo $[0, 1)$ en dos intervalos

$$I_0(p) = [0, 1 - p), \quad I_1(p) = [1 - p, 1) \quad (34)$$

y en la sucesión de intervalos

$$J_0(p) = [0, e^{-p}), \quad J_k(p) = \left[\sum_{j=0}^{k-1} e^{-p} \frac{p^j}{j!}, \sum_{j=0}^k e^{-p} \frac{p^j}{j!} \right), \quad k = 1, 2, \dots \quad (35)$$

Consideramos una variable aleatoria U con distribución $\mathcal{U}[0, 1)$ y construimos dos variables aleatorias V y W con distribuciones $\text{Bernoulli}(p)$ y $\text{Poisson}(p)$, respectivamente:

$$V := \mathbf{1}\{U \in I_1(p)\}, \quad W := \sum_{k=0}^{\infty} k \mathbf{1}\{U \in J_k(p)\}. \quad (36)$$

De la desigualdad $1 - p \leq e^{-p}$ resulta que $I_0(p) \subset J_0(p)$ y que $J_1(p) \subset I_1(p)$. En consecuencia, $V = W \iff U \in I_0(p) \cup J_1(p)$. Por ende,

$$\mathbb{P}(V = W) = \mathbb{P}(U \in I_0(p) \cup J_1(p)) = 1 - p + e^{-p}p, \quad (37)$$

y en consecuencia,

$$\mathbb{P}(V \neq W) = p - e^{-p}p = p(1 - e^{-p}) \leq p^2. \quad (38)$$

Usando la desigualdad (38) pueden obtenerse las siguientes cotas:

$$\sup_{k \geq 0} |\mathbb{P}(V = k) - \mathbb{P}(W = k)| \leq p^2, \quad (39)$$

$$\sum_k |\mathbb{P}(V = k) - \mathbb{P}(W = k)| \leq 2p^2. \quad (40)$$

La cota (39) se deduce de observar que

$$\begin{aligned} |\mathbb{P}(V = k) - \mathbb{P}(W = k)| &= |\mathbb{E}[\mathbf{1}\{V = k\}] - \mathbb{E}[\mathbf{1}\{W = k\}]| \\ &= |\mathbb{E}[\mathbf{1}\{V = k\} - \mathbf{1}\{W = k\}]| \\ &\leq \mathbb{E}[|\mathbf{1}\{V = k\} - \mathbf{1}\{W = k\}|] \\ &\leq \mathbb{E}[\mathbf{1}\{V \neq W\}] \\ &= \mathbb{P}(V \neq W). \end{aligned}$$

La cota (40) se deduce de observar que para todo $k = 0, 1, \dots$

$$\begin{aligned} |\mathbb{P}(V = k) - \mathbb{P}(W = k)| &= |\mathbb{P}(V = k, W \neq k) - \mathbb{P}(W = k, V \neq k)| \\ &\leq \mathbb{P}(V = k, V \neq W) + \mathbb{P}(W = k, V \neq W), \end{aligned}$$

y luego sumar sobre los posibles valores de k :

$$\sum_k |\mathbb{P}(V = k) - \mathbb{P}(W = k)| \leq 2\mathbb{P}(V \neq W).$$

□

Nota Bene. Esta técnica, denominada técnica de acoplamiento de variables aleatorias, permite probar (sin usar la fórmula de Stirling) que la distribución Binomial converge a la distribución Poisson.

Teorema 2.5 (Le Cam). *Sean X_1, \dots, X_n variables aleatorias independientes con distribución Bernoulli de parámetros p_1, \dots, p_n , respectivamente y sea $S = \sum_{i=1}^n X_i$. Entonces*

$$\sum_k |\mathbb{P}(S = k) - \mathbb{P}(N = k)| \leq 2 \sum_{i=1}^n p_i^2, \quad (41)$$

donde N es una variable aleatoria con distribución Poisson de media $\lambda = \sum_{i=1}^n p_i$.

Demostración. Sean U_1, \dots, U_n variables aleatorias independientes con distribución común $U[0, 1]$. Construimos variables aleatorias acopladas $V_i \sim \text{Bernoulli}(p_i)$ y $W_i \sim \text{Poisson}(p_i)$, $i = 1, \dots, n$:

$$V_i := \mathbf{1}\{U_i \in I_1(p_i)\}, \quad W_i := \sum_{k=0}^{\infty} k \mathbf{1}\{U_i \in J_k(p_i)\},$$

y las sumamos

$$S^* = \sum_{i=1}^n V_i, \quad N = \sum_{i=1}^n W_i.$$

Por construcción, las variables V_1, \dots, V_n son independientes y con distribución Bernoulli(p_i), respectivamente, y entonces, la variable S^* tiene la misma distribución que S ; las variables W_1, \dots, W_n son independientes y tienen distribución Poisson(p_i), respectivamente, y entonces, la variable N tiene distribución Poisson de media $\lambda = \sum_{i=1}^n p_i$.

Observando que cada k

$$|\mathbb{P}(S^* = k) - \mathbb{P}(N = k)| \leq \mathbb{P}(S^* = k, N \neq k) + \mathbb{P}(N = k, S^* \neq k).$$

se obtiene que

$$\sum_k |\mathbb{P}(S^* = k) - \mathbb{P}(N = k)| \leq 2\mathbb{P}(S^* \neq N).$$

Si $S^* \neq N$, entonces $V_i \neq W_i$ para algún $i = 1, \dots, n$. En consecuencia,

$$\mathbb{P}(S^* \neq N) \leq \sum_{i=1}^n \mathbb{P}(V_i \neq W_i) \leq \sum_{i=1}^n p_i^2.$$

□

Corolario 2.6 (Aproximación Poisson). *Para cada $k \geq 0$*

$$\lim_{n \rightarrow \infty} \binom{n}{k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \left(\frac{\lambda}{n}\right)^k = e^{-\lambda} \frac{\lambda^k}{k!}$$

Demostración. Sean U_1, \dots, U_n variables aleatorias independientes con distribución común $\mathcal{U}[0, 1)$. Para cada $i = 1, \dots, n$ definimos parejas de variables aleatorias (V_i, W_i) independientes

$$V_i := \mathbf{1}\{U_i \in I_1(p)\}, \quad W_i := \sum_{k=0}^{\infty} k \mathbf{1}\{U_i \in J_k(p)\}.$$

Por construcción, $V_i \sim \text{Bernoulli}(p)$ y $W_i \sim \text{Poisson}(p)$, en consecuencia las sumas

$$S = \sum_{i=1}^n V_i, \quad N = \sum_{i=1}^n W_i$$

son variables aleatorias con distribuciones Binomial(n, p) y Poisson(np), respectivamente. De acuerdo con la demostración del Teorema de Le Cam tenemos que

$$\left| \binom{n}{k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \left(\frac{\lambda}{n}\right)^k - e^{-\lambda} \frac{\lambda^k}{k!} \right| = |\mathbb{P}(S = k) - \mathbb{P}(N = k)| \leq 2np^2 = 2 \frac{\lambda^2}{n} \rightarrow 0.$$

□

Teorema 2.7. Supongamos que para cada n , $X_{n,1}, \dots, X_{n,r_n}$ son variables aleatorias independientes con distribución Bernoulli($p_{n,k}$). Si

$$\sum_{k=1}^{r_n} p_{n,k} \rightarrow \lambda \geq 0, \quad \max_{1 \leq k \leq r_n} p_{n,k} \rightarrow 0, \quad (42)$$

entonces

$$\mathbb{P}\left(\sum_{k=1}^{r_n} X_{n,k} = i\right) \rightarrow e^{-\lambda} \frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \dots \quad (43)$$

Si $\lambda = 0$, el límite (43) se interpreta como 1 para $i = 0$ y 0 para $i \geq 1$. En el caso $r_n = n$ y $p_{n,k} = \lambda/n$, (43) es la aproximación Poisson a la binomial. Notar que si $\lambda > 0$, entonces (42) implica que $r_n \rightarrow \infty$.

Demostración. Sea U_1, U_2, \dots una sucesión de variables aleatorias independientes, con distribución común $\mathcal{U}[0, 1)$. Definimos

$$V_{n,k} := \mathbf{1}\{U_k \in I_1(p_{n,k})\}.$$

Las variables $V_{n,1}, \dots, V_{n,r_n}$ son independientes y con distribución Bernoulli($p_{n,k}$). Puesto que $V_{n,1}, \dots, V_{n,r_n}$ tienen la misma distribución que $X_{n,1}, \dots, X_{n,r_n}$, (43) se obtiene mostrando que $V_n = \sum_{k=1}^{r_n} V_{n,k}$ satisface

$$\mathbb{P}(V_n = i) \rightarrow e^{-\lambda} \frac{\lambda^i}{i!}. \quad (44)$$

Ahora definimos

$$W_{n,k} := \sum_{i=0}^{\infty} i \mathbf{1}\{U_k \in J_i(p_{n,k})\}$$

$W_{n,k}$ tiene distribución Poisson de media $p_{n,k}$. Puesto que las $W_{n,k}$ son independientes, $W_n = \sum_{k=1}^{r_n} W_{n,k}$ tiene distribución Poisson de media $\lambda_n = \sum_{k=1}^{r_n} p_{n,k}$. De la desigualdad $1-p \leq e^{-p}$, se obtiene como consecuencia que

$$\begin{aligned}\mathbb{P}(V_{n,k} \neq W_{n,k}) &= \mathbb{P}(V_{n,k} = 1 \neq W_{n,k}) = \mathbb{P}(U_k \in I_1(p_{n,k}) - J_1(p_{n,k})) \\ &= p_{n,k} - e^{-p_{n,k}} p_{n,k} \leq p_{n,k}^2,\end{aligned}$$

y por (42)

$$\mathbb{P}(V_n \neq W_n) \leq \sum_{k=1}^{r_n} p_{n,k}^2 \leq \lambda_n \max_{1 \leq k \leq r_n} p_{n,k} \rightarrow 0.$$

(44) y (43) se obtienen de observar que

$$\mathbb{P}(W_n = i) = e^{-\lambda_n} \frac{\lambda_n^i}{i!} \rightarrow e^{-\lambda} \frac{\lambda^n}{n!}.$$

□

3. Cuentas con exponentiales

3.1. Motivación: pasaje de lo discreto a lo continuo

Para fijar ideas consideraremos una conversación telefónica y supondremos que su duración es un número entero de segundos. La duración de la conversación será tratada como una variable aleatoria T cuya distribución de probabilidades $p_n = \mathbb{P}(T = n)$ es conocida. La línea telefónica representa un sistema físico con dos estados posibles “ocupada” (E_0) y “libre” (E_1).

Imaginemos que cada segundo se decide si la conversación continúa o no por medio de una moneda cargada. En otras palabras, se realiza una sucesión de ensayos Bernoulli con probabilidad de éxito p a una tasa de un ensayo por segundo y se continúa hasta el primer éxito. La conversación termina cuando ocurre el primer éxito. En este caso la duración total de la conversación, el *tiempo de espera*, tiene distribución geométrica $p_n = (1-p)^{n-1}p$. Si en un instante cualquiera la línea está ocupada, la probabilidad que permanezca ocupada por más de un segundo es $(1-p)$, y la probabilidad de transición $E_0 \rightarrow E_1$ en el siguiente paso es p . En este caso esas probabilidades son independientes de cuánto tiempo estuvo ocupada la línea.

La descripción de los tiempos de espera mediante modelos discretos presupone la cuantización del tiempo y que los cambios solo pueden ocurrir en las épocas $\varepsilon, 2\varepsilon, \dots$. El tiempo de espera T más sencillo es el tiempo de espera hasta el primer éxito en una sucesión de ensayos Bernoulli con probabilidad de éxito $p(\varepsilon)$. En tal caso $\mathbb{P}(T > n\varepsilon) = (1 - p(\varepsilon))^n$ y el tiempo medio de espera es $\mathbb{E}[T] = \varepsilon/p(\varepsilon)$. Este modelo puede ser refinado haciendo que ε sea cada vez más chico pero manteniendo fija la esperanza $\varepsilon/p(\varepsilon) = 1/\lambda$. Para un intervalo de duración t corresponden aproximadamente $n \approx t/\varepsilon$ ensayos, y entonces para ε pequeño

$$\mathbb{P}(T > t) \approx (1 - \lambda\varepsilon)^{t/\varepsilon} \approx e^{-\lambda t}. \quad (45)$$

Este modelo considera el tiempo de espera como una variable aleatoria discreta distribuida geométricamente y (45) dice que “en el límite” se obtiene una distribución exponencial.

Si no discretizamos el tiempo tenemos que tratar con variables aleatorias continuas. El rol de la distribución geométrica para los tiempos de espera lo ocupa la *distribución exponencial*. Es la única variable continua dotada de una completa falta de memoria. En otras palabras, la probabilidad de que una conversación que llegó hasta el tiempo t continúe más allá del tiempo $t + s$ es independiente de la duración pasada de la conversación si, y solo si, la probabilidad que la conversación dure por lo menos t unidades de tiempo está dada por una exponencial $e^{-\lambda t}$.

Nota Bene Si en un momento arbitrario t la línea está ocupada, entonces la probabilidad de un cambio de estado durante el próximo segundo depende de cuan larga ha sido la conversación. En otras palabras, *el pasado influye sobre el futuro*. Esta circunstancia es la fuente de muchas dificultades en problemas más complicados.

3.2. Distribución exponencial

Se dice que la variable aleatoria T tiene *distribución exponencial de intensidad* $\lambda > 0$ y se denota $T \sim \text{Exp}(\lambda)$ si la función de distribución de T es de la forma

$$F_T(t) := \mathbb{P}(T \leq t) = (1 - e^{-\lambda t}) \mathbf{1}\{t \geq 0\}. \quad (46)$$

En tal caso T admite la siguiente función densidad de probabilidades

$$f_T(t) = \lambda e^{-\lambda t} \mathbf{1}\{t \geq 0\}. \quad (47)$$

Media y Varianza. Los valores de la esperanza y la varianza de T son, respectivamente, $\mathbb{E}[T] = 1/\lambda$ y $\mathbb{V}(T) = 1/\lambda^2$.

3.3. Suma de exponentiales independientes de igual intensidad

Teorema 3.1. Sean T_1, T_2, \dots, T_n variables aleatorias independientes, idénticamente distribuidas, con distribución exponencial de intensidad $\lambda > 0$. La suma $S_n = T_1 + \dots + T_n$ admite una densidad de probabilidades de la forma

$$f_{S_n}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \mathbf{1}\{t > 0\} \quad (48)$$

y su función de distribución es

$$F_{S_n}(t) = \left(1 - e^{-\lambda t} \sum_{i=0}^{n-1} \frac{(\lambda t)^i}{i!} \right) \mathbf{1}\{t \geq 0\}. \quad (49)$$

En otras palabras, la suma de n variables aleatorias independientes exponenciales de intensidad $\lambda > 0$ tiene distribución Gamma de parámetros n y λ : $\Gamma(n, \lambda)$.

Demostración. Por inducción. Para $n = 1$ no hay nada que probar: $S_1 = T_1 \sim \text{Exp}(\lambda)$. Supongamos ahora que la suma $S_n = T_1 + \dots + T_n$ admite una densidad de la forma (48). Debido a que las variables aleatorias S_n y T_{n+1} son independientes, la densidad de $S_{n+1} = S_n + T_{n+1}$ se obtiene convolucionando las densidades de S_n y T_{n+1} :

$$\begin{aligned} f_{S_{n+1}}(t) &= (f_{S_n} * f_{T_{n+1}})(t) = \int_0^t f_{S_n}(t-x) f_{T_{n+1}}(x) dx \\ &= \int_0^t \lambda e^{-\lambda(t-x)} \frac{(\lambda(t-x))^{n-1}}{(n-1)!} \lambda e^{-\lambda x} dx \\ &= \lambda e^{-\lambda t} \frac{\lambda^n}{(n-1)!} \int_0^t (t-x)^{n-1} dx = \lambda e^{-\lambda t} \frac{\lambda^n}{(n-1)!} \frac{t^n}{n} \\ &= \lambda e^{-\lambda t} \frac{(\lambda t)^n}{n!}. \end{aligned}$$

Las funciones de distribución (49) se obtienen integrando las densidades (48). Sea $t \geq 0$, integrando por partes puede verse que

$$\begin{aligned} F_{S_n}(t) &= \int_0^t f_{S_n}(s) ds = \int_0^t \frac{(\lambda s)^{n-1}}{(n-1)!} \lambda e^{-\lambda s} ds \\ &= -\frac{(\lambda s)^{n-1}}{(n-1)!} e^{-\lambda s} \Big|_0^t + \int_0^t \frac{(\lambda s)^{n-2}}{(n-2)!} \lambda e^{-\lambda t} ds \\ &= -\frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t} + F_{S_{n-1}}(t). \end{aligned} \tag{50}$$

Iterando (50) obtenemos (49). □

Nota Bene. En la demostración anterior se utilizó el siguiente resultado: si T_1, \dots, T_n son variables aleatorias independientes, entonces funciones (medibles) de familias disjuntas de las T_i también son independientes. (Para más detalles ver el Capítulo 1 de Durrett, R., (1996). *Probability Theory and Examples*, Duxbury Press, New York.) □

3.4. Mínimos

Lema 3.2. Sean T_1 y T_2 dos variables aleatorias independientes y exponenciales de intensidades λ_1 y λ_2 , respectivamente. Vale que

$$\mathbb{P}(T_1 < T_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \tag{51}$$

Demostración. La probabilidad $\mathbb{P}(T_1 < T_2)$ puede calcularse condicionando sobre T_1 :

$$\begin{aligned} \mathbb{P}(T_1 < T_2) &= \int_0^\infty \mathbb{P}(T_1 < T_2 | T_1 = t) f_{T_1}(t) dt = \int_0^\infty \mathbb{P}(t < T_2) \lambda_1 e^{-\lambda_1 t} dt \\ &= \lambda_1 \int_0^\infty e^{-\lambda_2 t} e^{-\lambda_1 t} dt = \lambda_1 \int_0^\infty e^{-(\lambda_1 + \lambda_2)t} dt = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \end{aligned}$$

□

Teorema 3.3. Sean T_1, T_2, \dots, T_n variables aleatorias exponenciales independientes de intensidades $\lambda_1, \lambda_2, \dots, \lambda_n$, respectivamente. Sean T y J las variables aleatorias definidas por

$$T := \min_i T_i, \quad J := \text{índice que realiza } T.$$

Entonces, T tiene distribución exponencial de intensidad $\lambda_1 + \dots + \lambda_n$ y

$$\mathbb{P}(J = j) = \frac{\lambda_j}{\lambda_1 + \dots + \lambda_n}.$$

Más aún, las variables T y J son independientes.

Demostración. En primer lugar, hay que observar que $T > t$ si y solo si $T_i > t$ para todo $i = 1, \dots, n$. Como las variables T_1, T_2, \dots, T_n son exponenciales independientes de intensidades $\lambda_1, \lambda_2, \dots, \lambda_n$ tenemos que

$$\mathbb{P}(T > t) = \prod_{i=1}^n \mathbb{P}(T_i > t) = \prod_{i=1}^n e^{-\lambda_i t} = e^{-(\lambda_1 + \dots + \lambda_n)t}.$$

Por lo tanto, T tiene distribución exponencial de intensidad $\lambda_1 + \dots + \lambda_n$.

En segundo lugar hay que observar que $J = j$ si y solo si $T = T_j$. Por lo tanto,

$$\mathbb{P}(J = j) = \mathbb{P}(T_j = \min_i T_i) = \mathbb{P}(T_j < \min_{i \neq j} T_i) = \frac{\lambda_j}{\lambda_1 + \dots + \lambda_n}.$$

La última igualdad se obtiene utilizando el Lema 3.2 pues las variables T_j y $\min_{i \neq j} T_i$ son independientes y exponenciales con intensidades λ_j y $\sum_{i \neq j} \lambda_i$, respectivamente.

Finalmente, si para cada j definimos $U_j = \min_{i \neq j} T_i$, tenemos que

$$\begin{aligned} \mathbb{P}(J = j, T \geq t) &= \mathbb{P}(t \leq T_j < U_j) \\ &= \int_t^\infty \mathbb{P}(T_j < U_j | T_j = s) \lambda_j e^{-\lambda_j s} ds \\ &= \lambda_j \int_t^\infty \mathbb{P}(U_j > s) e^{-\lambda_j s} ds = \lambda_j \int_t^\infty e^{-(\sum_{i \neq j} \lambda_i)s} e^{-\lambda_j s} ds \\ &= \frac{\lambda_j}{\lambda_1 + \dots + \lambda_n} \int_t^\infty (\lambda_1 + \dots + \lambda_n) e^{-(\lambda_1 + \dots + \lambda_n)s} ds \\ &= \frac{\lambda_j}{\lambda_1 + \dots + \lambda_n} e^{-(\lambda_1 + \dots + \lambda_n)t}. \end{aligned}$$

Lo que completa la demostración. □

Ejercicios adicionales

4. Sean T_1 y T_2 variables aleatorias independientes exponenciales de intensidad 2. Sean $T_{(1)} = \min(T_1, T_2)$ y $T_{(2)} = \max(T_1, T_2)$. Hallar la esperanza y la varianza de $T_{(1)}$ y de $T_{(2)}$.

- 5.** *Suma geométrica de exponentiales independientes.* Sean T_1, T_2, \dots variables aleatorias independientes idénticamente distribuidas con ley exponencial de intensidad λ . Se define $T = \sum_{i=1}^N T_i$, donde N es una variable aleatoria con distribución geométrica de parámetro p , independiente de las variables T_1, T_2, \dots . Hallar la distribución de T . (*Sugerencia:* Utilizar la fórmula de probabilidad total condicionando a los posibles valores de N y el desarrollo en serie de Taylor de la función exponencial.)
-

4. Bibliografía consultada

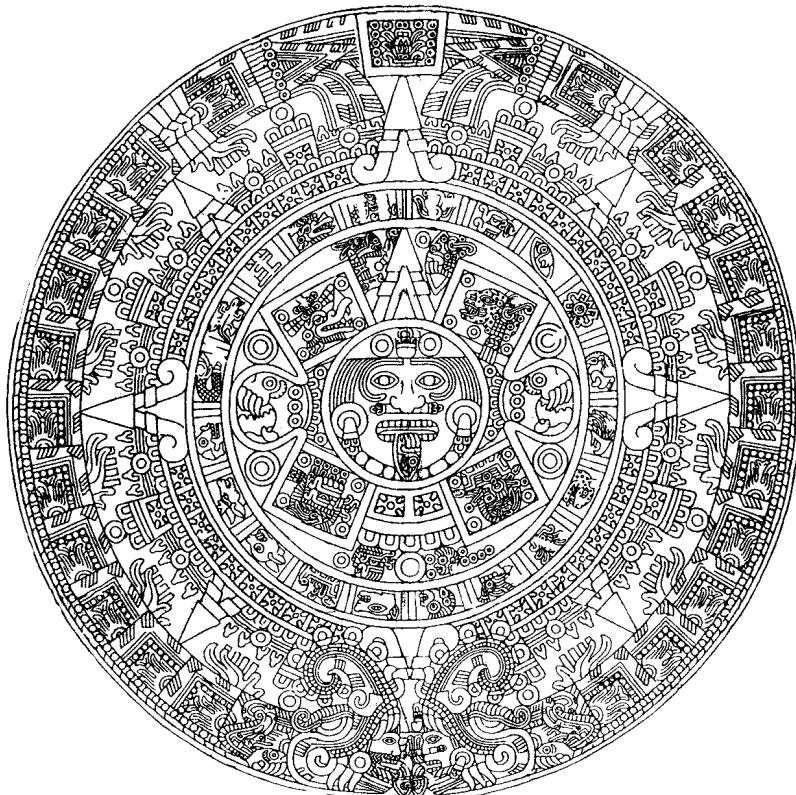
Para redactar estas notas se consultaron los siguientes libros:

1. Billingsley, P.: Probability and measure. John Wiley & Sons, New York. (1986)
2. Durrett R.: Probability. Theory and Examples. Duxbury Press, Belmont. (1996)
3. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons, New York. (1957)
4. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 2. John Wiley & Sons, New York. (1971)
5. Grimmett, G. R., Stirzaker, D. R.: Probability and Random Processes. Oxford University Press, New York. (2001)
6. Meester, R.: A Natural Introduction to Probability Theory. Birkhauser, Berlin. (2008).
7. Meyer, P. L.: Introductory Probability and Statistical Applications. Addison-Wesley, Massachusetts. (1972)
8. Ross, S. M: Introduction to Probability and Statistics for Engineers and Scientists. Elsevier Academic Press, San Diego. (2004)
9. Soong, T. T.: Fundamentals of Probability and Statistics for Engineers. John Wiley & Sons Ltd. (2004)

Procesos de Poisson
(Borradores, Curso 23)

Sebastian Grynberg

22 de abril de 2013



ollin tonatiuh

*el tiempo sólo es tardanza
de lo que está por venir*

(Martín Fierro)

Índice

1. Proceso puntual de Poisson	2
1.1. Procesos puntuales	2
1.2. Procesos de Poisson	4
1.3. Construcción	5
1.4. Distribución condicional de los tiempos de llegada	10
1.5. Coloración y adelgazamiento de procesos de Poisson	11
1.6. Superposición de Procesos de Poisson: competencia	13
1.7. Procesos de Poisson compuestos	15
2. Bibliografía consultada	17

1. Proceso puntual de Poisson

1.1. Procesos puntuales

Informalmente, un proceso puntual aleatorio es un conjunto enumerable de puntos aleatorios ubicados sobre la recta real. En la mayoría de las aplicaciones un *punto* de un proceso puntual es el instante en que ocurre algún evento, motivo por el cual los puntos también se llaman *eventos* o *arribos*. Por ejemplo, los tiempos de arribo de clientes a la caja de un supermercado o de los trabajos al procesador central de una computadora son procesos puntuales. En teoría fiabilidad, un evento podría ser el instante en que ocurre una falla. El ejemplo básico de este tipo de procesos es el *proceso de Poisson*.

Definición 1.1 (Proceso puntual aleatorio). *Un proceso puntual aleatorio sobre la semi-recta positiva es una sucesión $\{S_n : n \geq 0\}$ de variables aleatorias no negativas tales que, casi seguramente,*

- (a) $S_0 \equiv 0$,
- (b) $0 < S_1 < S_2 < \dots$,
- (c) $\lim_{n \rightarrow \infty} S_n = +\infty$.

La condición (b) significa que no hay arribos simultáneos. La condición (c) significa que no hay *explosiones*, esto es, no hay una acumulación de arribos en tiempos finitos.

La sucesión de variables aleatorias $\{T_n : n \geq 1\}$ definida por

$$T_n := S_n - S_{n-1} \quad (1)$$

se llama la *sucesión de tiempos de espera entre arribos*.

Introducimos una familia de nuevas variables aleatorias $N(t)$, $t \geq 0$, de la siguiente manera: para cada $t \geq 0$ definimos $N(t)$ como la cantidad de arribos ocurridos durante el intervalo de tiempo $(0, t]$,

$$N(t) := \sum_{n \geq 1} \mathbf{1}\{S_n \leq t\} \quad (2)$$

$$= \max\{n \geq 0 : S_n \leq t\}. \quad (3)$$

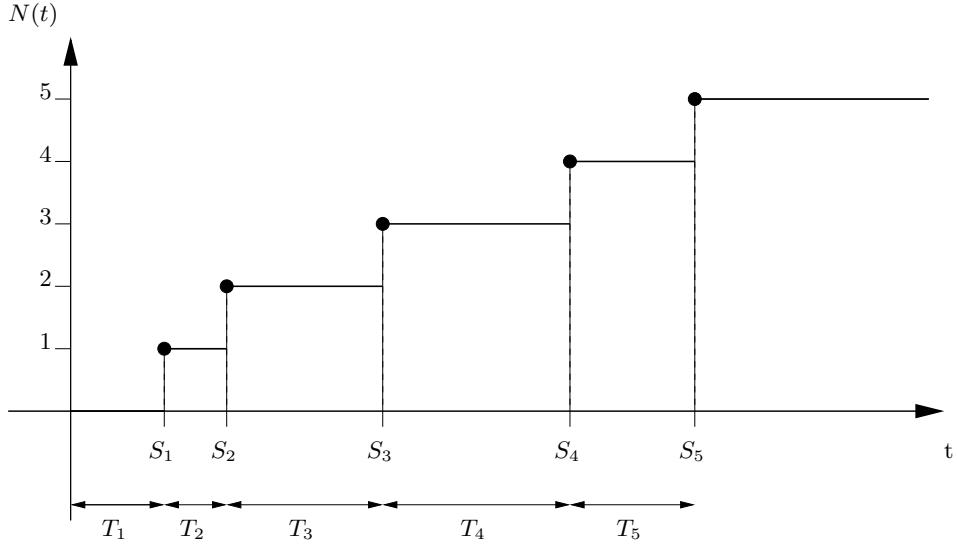


Figura 1: Realización típica de un proceso puntual aleatorio sobre la semi-recta positiva.

Observación 1.2. Notar que $N(t)$ es una función de t y de las variables aleatorias T_1, T_2, \dots a valores enteros no negativos. Indicaremos esa relación de la siguiente manera

$$N(t) = \Psi(t; T_1, T_2, \dots), \quad (4)$$

donde Ψ es la relación definida en (2).

La cantidad de arribos ocurridos durante el intervalo de tiempo $(s, t] \subset \mathbb{R}^+$, $N(s, t]$, es el incremento $N(t) - N(s)$

$$N(s, t] := N(t) - N(s) = \sum_{n \geq 1} \mathbf{1}\{s < S_n \leq t\}. \quad (5)$$

De (3) se obtiene la relación básica que conecta a las variables $N(t)$ con las S_n :

$$N(t) \geq n \iff S_n \leq t. \quad (6)$$

De allí se desprende que

$$N(t) = n \iff S_n \leq t < S_{n+1}. \quad (7)$$

Proceso de conteo. La familia de variables aleatorias $\{N(t) : t \geq 0\}$ es un proceso estocástico denominado el *proceso de conteo* de la sucesión de arribos $\{S_n : n \geq 0\}$. Debido a que la sucesión de arribos se puede reconstruir a partir de N , N también recibe la denominación “*proceso puntual*”.

Propiedades. Por definición, el proceso de conteo satisface las siguientes propiedades:

- (i) Para cada $t \geq 0$, la variable aleatoria $N(t)$ tiene valores enteros no negativos.
- (ii) $N(0) = 0$ y $\lim_{t \rightarrow \infty} N(t) = \infty$.

(iii) Si $s < t$, entonces $N(s) \leq N(t)$.

(iv) Como el intervalo $(0, t]$ es cerrado a la derecha, la función (aleatoria) $N : \mathbb{R}^+ \rightarrow \mathbb{N}_0$ es continua a derecha. Además, en los puntos de discontinuidad tiene saltos de longitud 1.

En otras palabras, el gráfico de la función aleatoria $N : \mathbb{R}^+ \rightarrow \mathbb{N}_0$ es una escalera no decreciente, continua a derecha y con saltos de longitud 1 en cada uno de los arribos del proceso puntual.

Programa. En lo que sigue estudiaremos la distribución conjunta de las $N(t)$ bajo ciertas condiciones sobre los tiempos de espera entre arribos T_n y vice versa.

1.2. Procesos de Poisson

Existen varias definiciones equivalentes de procesos de Poisson. Adoptamos la que nos parece más sencilla y generalizable.¹

Definición 1.3 (Proceso de Poisson). *Un proceso puntual $\{S_n : n \geq 0\}$ sobre la semi-recta positiva es un proceso de Poisson de intensidad $\lambda > 0$ si satisface las siguientes condiciones*

- (i) El proceso tiene *incrementos independientes*: para cada colección finita de tiempos $0 = t_0 < t_1 < \dots < t_n$, los incrementos $N(t_{i-1}, t_i] = N(t_i) - N(t_{i-1})$, $i = 1, \dots, n$ son independientes.
- (ii) Los *incrementos individuales* $N(s, t] = N(t) - N(s)$ tienen la distribución Poisson:

$$\mathbb{P}(N(s, t] = n) = e^{-\lambda(t-s)} \frac{(\lambda(t-s))^n}{n!}, \quad n = 0, 1, \dots, 0 \leq s < t. \quad (8)$$

Nota Bene. La condición (ii) de la Definición 1.3 se puede descomponer en dos partes.

(a) *Los incrementos son temporalmente homogéneos* (i.e., la distribución de los incrementos depende solamente de la longitud del intervalo de tiempo pero no de su posición) y (b) *la distribución de cada incremento individual es Poisson de media proporcional a la cantidad de tiempo considerado*.

Que un proceso puntual sea *temporalmente homogéneo* y que tenga *incrementos independientes* significa que si se lo reinicia desde cualquier instante de tiempo t , el proceso así obtenido es independiente de todo lo que ocurrió previamente (por tener incrementos independientes) y que tiene la misma distribución que el proceso original (por ser temporalmente homogéneo). En otras palabras, el proceso no tiene memoria.

Es de suponer que, bajo esas condiciones, los tiempos de espera entre arribos tienen que ser variables aleatorias independientes, cada una con distribución exponencial del mismo parámetro. Ésto último es consistente con la condición sobre la distribución que tienen los incrementos individuales (8).

¹Elegimos la Definición 1.3 porque tiene la virtud de que se puede extender a \mathbb{R}^d sin ninguna dificultad: un subconjunto aleatorio (numerable) Π de \mathbb{R}^d se llama un *proceso de Poisson de intensidad λ* si, para todo $A \in \mathcal{B}(\mathbb{R}^d)$, las variables aleatorias $N(A) = |\Pi \cap A|$ satisfacen (a) $N(A)$ tiene la distribución Poisson de parámetro $\lambda|A|$, y (b) Si $A_1, A_2, \dots, A_n \in \mathcal{B}(\mathbb{R}^d)$ son conjuntos disjuntos, entonces $N(A_1), N(A_2), \dots, N(A_n)$ son variables aleatorias independientes.

En efecto, de la relación básica (6) se deduce que si $\{S_n : n \geq 0\}$ es un proceso de Poisson de intensidad λ , entonces las variables S_n tienen distribución $\Gamma(n, \lambda)$:

$$\mathbb{P}(S_n > t) = \mathbb{P}(N(t) < n) = \sum_{k=0}^{n-1} \mathbb{P}(N(t) = k) = \sum_{k=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^k}{k!}.$$

□

1.3. Construcción

En lo que sigue mostraremos una forma de construir un proceso puntual de Poisson $\{S_n : n \geq 0\}$ de intensidad λ . Los arribos, S_n , se construyen utilizando una sucesión de variables aleatorias a valores positivos $\{T_n : n \geq 1\}$:

$$S_0 := 0, \quad S_n := \sum_{i=1}^n T_i, \quad n = 1, 2, \dots \quad (9)$$

Teorema 1.4. *Sea $\{T_n : n \geq 1\}$ una sucesión de variables aleatorias independientes, cada una con distribución exponencial de intensidad λ . El proceso de arribos $\{S_n : n \geq 0\}$ definido en (9) es un proceso puntual de Poisson de intensidad λ .* (Ver la Definición 1.3).

Demostración.

1. Proceso Puntual. Para cada $n \geq 1$, $\mathbb{P}(T_n > 0) = 1$ y por la ley fuerte de los grandes números $\frac{1}{n} \sum_{i=1}^n T_i \rightarrow \frac{1}{\lambda}$ casi seguramente. Por lo tanto, $\{S_n : n \geq 0\}$ es un proceso puntual.

2. Distribuciones Poisson. Para cada $n \geq 1$, $S_n = T_1 + \dots + T_n$ tiene distribución $\Gamma(n, \lambda)$:

$$F_{S_n}(t) = \mathbb{P}(S_n \leq t) = \left(1 - e^{-\lambda t} \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} \right) \mathbf{1}\{t \geq 0\} = \left(e^{-\lambda t} \sum_{k=n}^{\infty} \frac{(\lambda t)^k}{k!} \right) \mathbf{1}\{t \geq 0\}.$$

Observando que $\{N(t) = n\} = \{N(t) < n+1\} \setminus \{N(t) < n\}$ y usando la relación básica, $N(t) < n \iff S_n > t$, se deduce que

$$\begin{aligned} \mathbb{P}(N(t) = n) &= \mathbb{P}(N(t) < n+1) - \mathbb{P}(N(t) < n) = \mathbb{P}(S_{n+1} > t) - \mathbb{P}(S_n > t) \\ &= e^{-\lambda t} \sum_{k=0}^n \frac{(\lambda t)^k}{k!} - e^{-\lambda t} \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad n = 0, 1, \dots \end{aligned} \quad (10)$$

Por lo tanto, para cada $t > 0$ fijo, el incremento $N(t)$ tiene una distribución Poisson de media λt :

$$N(t) \sim \text{Poisson}(\lambda t).$$

3. Pérdida de memoria. Fijamos $t > 0$ y consideramos los arribos posteriores al instante t . Por (3) tenemos que $S_{N(t)} \leq t < S_{N(t)+1}$. El tiempo de espera desde t hasta el primer arribo posterior a t es $S_{N(t)+1} - t$; el tiempo de espera entre el primer y el segundo arribo posteriores a t es $T_{N(t)+2}$; y así siguiendo. De este modo

$$T_1^{(t)} := S_{N(t)+1} - t, \quad T_2^{(t)} := T_{N(t)+2}, \quad T_3^{(t)} := T_{N(t)+3}, \dots \quad (11)$$

definen los tiempos de espera entre arribos posteriores a t .

Debido a la independencia de las T_k y la propiedad de pérdida de memoria de la distribución exponencial, parece intuitivamente claro que condicionando al evento $\{N(t) = n\}$ las variables aleatorias (11) son independientes y con distribución exponencial.

En lo que sigue mostraremos que $N(t), T_1^{(t)}, T_2^{(t)}, \dots$ son variables aleatorias independientes y que

$$(T_1^{(t)}, T_2^{(t)}, \dots) \sim (T_1, T_2, \dots). \quad (12)$$

Basta mostrar que para todo $n \geq 0$ y para toda elección de números positivos t_1, \dots, t_m , $m \in \mathbb{N}$, vale que

$$\mathbb{P}(N(t) = n, T_1^{(t)} > t_1, \dots, T_m^{(t)} > t_m) = \mathbb{P}(N(t) = n) e^{-\lambda t_1} \cdots e^{-\lambda t_m}. \quad (13)$$

Para probarlo condicionaremos sobre la variable S_n ,

$$\begin{aligned} \mathbb{P}(N(t) = n, T_1^{(t)} > t_1) &= \mathbb{P}(S_n \leq t < S_{n+1}, S_{n+1} - t > t_1) \\ &= \mathbb{P}(S_n \leq t, T_{n+1} > t_1 + t - S_n) \\ &= \int_0^t \mathbb{P}(T_{n+1} > t_1 + t - s) f_{S_n}(s) ds \\ &= e^{-\lambda t_1} \int_0^t \mathbb{P}(T_{n+1} > t - s) f_{S_n}(s) ds \\ &= e^{-\lambda t_1} \mathbb{P}(S_n \leq t, T_{n+1} > t - S_n) \\ &= \mathbb{P}(N(t) = n) e^{-\lambda t_1}. \end{aligned}$$

Para obtener la segunda igualdad hay que observar que $\{S_{n+1} > t\} \cap \{S_{n+1} - t > t_1\} = \{S_{n+1} > t_1 + t\}$ y escribir $S_{n+1} = S_n + T_{n+1}$; la tercera se obtiene condicionando sobre S_n ; la cuarta se obtiene usando la propiedad de pérdida de memoria de la exponencial ($\mathbb{P}(T_{n+1} > t_1 + t - s) = \mathbb{P}(T_{n+1} > t_1) \mathbb{P}(T_{n+1} > t - s) = e^{-\lambda t_1} \mathbb{P}(T_{n+1} > t - s)$).

Por la independencia de las variables T_n ,

$$\begin{aligned} &\mathbb{P}(N(t) = n, T_1^{(t)} > t_1, \dots, T_m^{(t)} > t_m) \\ &= \mathbb{P}(S_n \leq t < S_{n+1}, S_{n+1} - t > t_1, T_{n+2} > t_2, T_{n+m} > t_m) \\ &= \mathbb{P}(S_n \leq t < S_{n+1}, S_{n+1} - t > t_1) e^{-\lambda t_2} \cdots e^{-\lambda t_m} \\ &= \mathbb{P}(N(t) = n) e^{-\lambda t_1} \cdots e^{-\lambda t_m}. \end{aligned}$$

4. Incrementos estacionarios e independientes. Por (6), $N(t+s) - N(t) \geq m$, o $N(t+s) \geq N(t) + m$, si y solo si $S_{N(t)+m} \leq t+s$, que es la misma cosa que $T_1^{(t)} + \cdots + T_m^{(t)} \leq s$. Así

$$N(t+s) - N(t) = \max\{m : T_1^{(t)} + \cdots + T_m^{(t)} \leq s\}. \quad (14)$$

Comparando (14) y (3) se puede ver que para t fijo las variables aleatorias $N(t+s) - N(t)$ para $s \geq 0$ se definen en términos de la sucesión (11) exactamente de la misma manera en que las $N(s)$ se definen en términos de la sucesión original de tiempos de espera. En otras palabras,

$$N(t+s) - N(t) = \Psi(s; T_1^{(t)}, T_2^{(t)}, \dots), \quad (15)$$

donde Ψ es la función definida en la Observación 4. De acuerdo con (12)

$$\{N(t+s) - N(t) : s \geq 0\} \sim \{N(s) : s \geq 0\}. \quad (16)$$

De (15) y lo visto en **3.** se deduce que $N(t)$ y $\{N(t+s) - N(t) : s \geq 0\}$ son independientes.

Sean $n \geq 2$ y $0 < t_1 < t_2 < \dots < t_n$. Como $(N(t_2) - N(t_1), \dots, N(t_n) - N(t_{n-1}))$ es una función de $\{N(t_1 + s) - N(t_1) : s \geq 0\}$, tenemos que

$$N(t_1) \text{ y } (N(t_2) - N(t_1), \dots, N(t_n) - N(t_{n-1}))$$

son independientes. Esto es,

$$\begin{aligned} & \mathbb{P}(N(t_1) = m_1, N(t_2) - N(t_1) = m_2, \dots, N(t_n) - N(t_{n-1}) = m_n) \\ &= \mathbb{P}(N(t_1) = m_1) \mathbb{P}(N(t_2) - N(t_1) = m_2, \dots, N(t_n) - N(t_{n-1}) = m_n) \end{aligned}$$

En particular, se obtiene la la independencia de los incrementos para el caso en que $n = 2$:

$$\mathbb{P}(N(t_1) = m_1, N(t_2) - N(t_1) = m_2) = \mathbb{P}(N(t_1) = m_1) \mathbb{P}(N(t_2) - N(t_1) = m_2).$$

Usando (16) se concluye que

$$\begin{aligned} & (N(t_2) - N(t_1), N(t_3) - N(t_2), \dots, N(t_n) - N(t_{n-1})) \\ & \sim (N(t_2 - t_1), N(t_3 - t_1) - N(t_2 - t_1), \dots, N(t_n - t_1) - N(t_{n-1} - t_1)). \end{aligned} \quad (17)$$

El caso general se obtiene por iteración del mismo argumento, aplicado al lado derecho de (17):

$$\begin{aligned} & \mathbb{P}(N(t_2) - N(t_1) = m_2, N(t_k) - N(t_{k-1}) = m_k, 3 \leq k \leq n) \\ &= \mathbb{P}(N(t_2 - t_1) = m_2, N(t_k - t_1) - N(t_{k-1} - t_1) = m_k, 3 \leq k \leq n) \\ &= \mathbb{P}(N(t_2 - t_1) = m_2) \mathbb{P}(N(t_k - t_1) - N(t_{k-1} - t_1) = m_k, 3 \leq k \leq n) \\ &= \mathbb{P}(N(t_2) - N(t_1) = m_2) \mathbb{P}(N(t_k) - N(t_{k-1}) = m_k, 3 \leq k \leq n) \\ &= \dots \\ &= \prod_{k=2}^n \mathbb{P}(N(t_k) - N(t_{k-1}) = m_k). \end{aligned}$$

Por lo tanto, si $0 = t_0 < t_1 < \dots < t_n$, entonces

$$\mathbb{P}(N(t_k) - N(t_{k-1}) = m_k, 1 \leq k \leq n) = \prod_{k=1}^n \mathbb{P}(N(t_k - t_{k-1}) = m_k). \quad (18)$$

De (18) y (10) se obtienen las dos condiciones que definen a un proceso de Poisson. \square

En lo que sigue mostraremos que vale la recíproca. Esto es, los tiempos de espera entre arribos de un proceso de Poisson de intensidad λ son variables aleatorias independientes cada una con distribución exponencial de intensidad λ .

Teorema 1.5. *Sea $\{S_n : n \geq 0\}$ un proceso puntual de Poisson de intensidad λ sobre la semi-recta positiva. Los tiempos de espera entre arribos T_n , $n \geq 1$, definidos en (1), constituyen una sucesión de variables aleatorias independientes cada una con distribución exponencial de intensidad λ .*

Demostración. La densidad conjunta de $\mathbf{T} = (T_1, T_2, \dots, T_n)$ se obtendrá a partir de la densidad conjunta de las variables $\mathbf{S} = (S_1, S_2, \dots, S_n)$ usando el método del Jacobiano. Por definición,

$$(T_1, T_2, \dots, T_n) = g(S_1, S_2, \dots, S_n),$$

donde $g : G_0 \rightarrow G$ es la transformación lineal biyectiva entre los conjuntos abiertos $G_0 = \{(s_1, \dots, s_n) \in \mathbb{R}^n : 0 < s_1 < s_2 < \dots < s_n\}$ y $G = \{(t_1, \dots, t_n) : t_1 > 0, \dots, t_n > 0\}$ definida por

$$g(s_1, s_2, \dots, s_n) = (s_1, s_2 - s_1, \dots, s_n - s_{n-1}).$$

La función inversa $h = g^{-1}$ es de la forma

$$h(t_1, \dots, t_n) = (t_1, t_1 + t_2, \dots, t_1 + \dots + t_n)$$

y sus derivadas parciales

$$\frac{\partial s_i}{\partial t_j} = \frac{\partial \sum_{k=1}^i t_k}{\partial t_j} = \mathbf{1}\{j \leq i\}, \quad 1 \leq i, j \leq n$$

son continuas en G . El jacobiano es

$$J(\mathbf{s}, \mathbf{t}) = \left| \left(\frac{\partial s_i}{\partial t_j} \right) \right| = 1$$

debido a que se trata de una matriz triangular inferior con 1's en la diagonal. Bajo esas condiciones tenemos que

$$f_{\mathbf{T}}(\mathbf{t}) = f_{\mathbf{S}}(h(\mathbf{t})) \mathbf{1}\{\mathbf{t} \in G\}.$$

La densidad conjunta de las variables (S_1, \dots, S_n) queda únicamente determinada por la relación

$$\mathbb{P}(\mathbf{S} \in A) = \int_A f_{\mathbf{S}}(\mathbf{s}) d\mathbf{s}, \quad A = (a_1, b_1] \times \dots \times (a_n, b_n] \subset G_0.$$

Supongamos que $0 = b_0 \leq a_1 < b_1 < a_2 < b_2 < \dots < a_n < b_n$ y calculemos la probabilidad del evento $\bigcap_{i=1}^n \{a_i < S_i \leq b_i\}$. Para ello observamos que $\bigcap_{i=1}^n \{a_i < S_i \leq b_i\} = \bigcap_{i=1}^{n-1} \{N(a_i) - N(b_{i-1}) = 0, N(b_i) - N(a_i) = 1\} \cap \{N(a_n) - N(b_{n-1}) = 0, N(b_n) - N(a_n) \geq 1\}$ y usamos las propiedades de independencia y homogeneidad temporal que caracterizan a los incrementos de un proceso de Poisson de intensidad λ :

$$\begin{aligned} & \mathbb{P} \left(\bigcap_{i=1}^n \{a_i < S_i \leq b_i\} \right) \\ &= \left(\prod_{i=1}^{n-1} e^{-\lambda(a_i - b_{i-1})} \lambda(b_i - a_i) e^{-\lambda(b_i - a_i)} \right) e^{-\lambda(a_n - b_{n-1})} (1 - e^{-\lambda(b_n - a_n)}) \\ &= \left(\prod_{i=1}^{n-1} \lambda(b_i - a_i) \right) e^{-\lambda a_n} (1 - e^{-\lambda(b_n - a_n)}) \\ &= \left(\prod_{i=1}^{n-1} \lambda(b_i - a_i) \right) (e^{-\lambda a_n} - e^{-\lambda b_n}) \\ &= \int_{a_1}^{b_1} \lambda ds_1 \cdots \int_{a_{n-1}}^{b_{n-1}} \lambda ds_{n-1} \int_{a_n}^{b_n} \lambda e^{-\lambda s_n} ds_n \\ &= \int_{a_1}^{b_1} \cdots \int_{a_{n-1}}^{b_{n-1}} \int_{a_n}^{b_n} \lambda^n e^{-\lambda s_n} ds_1 \cdots ds_{n-1} ds_n \end{aligned} \tag{19}$$

De (19) se deduce que la densidad conjunta de (S_1, \dots, S_n) es

$$f_{(S_1, \dots, S_n)}(s_1, \dots, s_n) = \lambda^n e^{-\lambda s_n} \mathbf{1}\{0 < s_1 < \dots < s_n\}.$$

Por lo tanto,

$$\begin{aligned} f_{(T_1, \dots, T_n)}(t_1, \dots, t_n) &= \lambda^n e^{-\lambda \sum_{i=1}^n t_i} \mathbf{1}\{t_1 > 0, \dots, t_n > 0\} \\ &= \prod_{i=1}^n \lambda e^{-\lambda t_i} \mathbf{1}\{t_i > 0\}. \end{aligned} \quad (20)$$

La identidad (20) significa que los tiempos de espera entre arribos son independientes cada uno con distribución exponencial de intensidad λ . \square

Ejemplo 1.6. Suponga que el flujo de inmigración de personas hacia un territorio es un proceso de Poisson de tasa $\lambda = 1$ por día.

- (a) ¿Cuál es el tiempo esperado hasta que se produce el arribo del décimo inmigrante?
- (b) ¿Cuál es la probabilidad de que el tiempo de espera entre el décimo y el undécimo arribo supere los dos días?

Solución:

- (a) $\mathbb{E}[S_{10}] = \frac{10}{\lambda} = 10$ días.
- (b) $\mathbb{P}(T_{11} > 2) = e^{-2\lambda} = e^{-2} \approx 0.133$.

\square

Ejercicios adicionales

1. En un sistema electrónico se producen fallas de acuerdo con un proceso de Poisson de tasa 2.5 por mes. Por motivos de seguridad se ha decidido cambiarlo cuando ocurran 196 fallas. Hallar la media y la varianza del tiempo de uso del sistema.

2. Sean T una variable aleatoria con distribución exponencial de media 2 y $\{N(t), t \geq 0\}$ un proceso de Poisson de tasa 10 (independiente de T). Hallar $\text{Cov}(T, N(T))$.

3. Sea $A(t) = t - S_{N(t)}$ el tiempo reverso al evento más reciente en un proceso de Poisson y sea $B(t) = S_{N(t)+1} - t$ el tiempo directo hasta el próximo evento. Mostrar que

- (a) $A(t)$ y $B(t)$ son independientes,
- (b) $B(t)$ se distribuye como T_1 (exponencial de intensidad λ) ,
- (c) $A(t)$ se distribuye como $\min(T_1, t)$:

$$\mathbb{P}(A(t) \leq x) = (1 - e^{-\lambda x}) \mathbf{1}\{0 \leq x < t\} + \mathbf{1}\{x \geq t\}.$$

4. Sea $L(t) = A(t) + B(t) = S_{N(t)+1} - S_{N(t)}$ la longitud del intervalo de tiempo entre arribos que contiene a t .

(a) Mostrar que $L(t)$ tiene densidad

$$d_t(x) = \lambda^2 x e^{-\lambda x} \mathbf{1}\{0 < x < t\} + \lambda(1 + \lambda t)e^{-\lambda x} \mathbf{1}\{x \geq t\}.$$

(b) Mostrar que $\mathbb{E}[L(t)]$ converge a $2\mathbb{E}[T_1]$ cuando $t \rightarrow \infty$. Esto parece una paradoja debido a que $L(t)$ es uno de los T_n . Dar una resolución intuitiva de esta paradoja.

1.4. Distribución condicional de los tiempos de llegada

Supongamos que sabemos que ocurrió exactamente un arribo de un proceso de Poisson en el intervalo $[0, t]$. Queremos determinar la distribución del tiempo en que el arribo ocurrió. Como el proceso de Poisson es temporalmente homogéneo y tiene incrementos independientes es razonable pensar que los intervalos de igual longitud contenidos en el intervalo $[0, t]$ deben tener la misma probabilidad de contener al arribo. En otras palabras, el tiempo en que ocurrió el arribo debe estar distribuido uniformemente sobre el intervalo $[0, t]$. Esto es fácil de verificar puesto que, para $s \leq t$,

$$\begin{aligned} \mathbb{P}(T_1 < s | N(t) = 1) &= \frac{\mathbb{P}(T_1 < s, N(t) = 1)}{\mathbb{P}(N(t) = 1)} \\ &= \frac{\mathbb{P}(1 \text{ arribo en } (0, s], 0 \text{ arribos en } (s, t])}{\mathbb{P}(N(t) = 1)} \\ &= \frac{\mathbb{P}(1 \text{ arribo en } (0, s]) \mathbb{P}(0 \text{ arribos en } (s, t])}{\mathbb{P}(N(t) = 1)} \\ &= \frac{\lambda s e^{-\lambda s} e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \\ &= \frac{s}{t} \end{aligned}$$

Este resultado puede generalizarse

Teorema 1.7 (Propiedad condicional). *Sea Π un proceso de Poisson de intensidad λ sobre \mathbb{R}^+ . Condicional al evento $N(t) = n$, los n arribos ocurridos en el intervalo $[0, t]$ tienen la misma distribución conjunta que la de n puntos independientes elegidos al azar sobre el intervalo $[0, t]$. En otras palabras, condicional a $N(t) = n$ los puntos en cuestión se distribuyen como n variables aleatorias independientes, cada una con distribución uniforme sobre el intervalo $[0, t]$.*

Demostración. Sea A_1, A_2, \dots, A_k una partición del intervalo $[0, t]$. Si $n_1 + n_2 + \dots + n_k = n$, entonces

$$\begin{aligned} \mathbb{P}(N(A_i) = n_i, 1 \leq i \leq k | N(t) = n) &= \frac{\prod_i \mathbb{P}(N(A_i) = n_i)}{\mathbb{P}(N(t) = n)} \\ &= \frac{\prod_i e^{-\lambda|A_i|} (\lambda|A_i|)^{n_i} / n_i!}{e^{-\lambda t} (\lambda t)^n / n!} \\ &= \frac{n!}{n_1! n_2! \cdots n_k!} \prod_i \left(\frac{|A_i|}{t} \right)^{n_i}. \end{aligned} \quad (21)$$

Por una parte la distribución condicional de las posiciones de los n arribos queda completamente caracterizada por esta función de A_1, \dots, A_k .

Por otra parte la distribución multinomial (21) es la distribución conjunta de n puntos independientes elegidos al azar de acuerdo con la distribución uniforme sobre el intervalo $[0, t]$.

En efecto, basta observar que si U_1, \dots, U_n son variables aleatorias independientes con distribución uniforme sobre un conjunto A , y $M(B) = \sum_i \mathbf{1}\{U_i \in B\}$, entonces

$$\mathbb{P}(M(B_i) = n_i, i = 1, \dots, k) = \frac{n!}{n_1! \cdots n_k!} \prod_{i=1}^k \left(\frac{|B_i|}{|A|} \right)^{n_i}.$$

Se infiere que la distribución conjunta de los puntos en $\Pi \cap [0, t]$ condicional a que hay exactamente n de ellos, es la misma que la de n puntos independientes elegidos al azar con la distribución uniforme sobre el intervalo $[0, t]$. \square

Nota Bene. La propiedad condicional permite probar la existencia de procesos de Poisson mediante simulación. Sea $\lambda > 0$ y sea A_1, A_2, \dots una partición de \mathbb{R}^d en conjuntos boreelianos de medida de Lebesgue finita. Para cada i , simulamos una variable aleatoria N_i con distribución Poisson de parámetro $\lambda |A_i|$. Luego muestreamos n puntos elegidos independientemente sobre A_i , cada uno con distribución uniforme sobre A_i . La unión sobre i de tales conjuntos de puntos es un proceso de Poisson de intensidad λ . (Para más detalles ver el Chap 7 de Ferrari, Galves (2001)) \square

Ejemplo 1.8 (Insectos en un asado). Todo tipo de insectos aterrizan en la mesa de un asado a la manera de un proceso de Poisson de tasa 3 por minuto. Si entre las 13:30 y las 13:35 aterrizaron 8 insectos, cuál es la probabilidad de que exactamente 3 de ellos hayan aterrizado durante el primer minuto?

Solución: Dado que aterrizaron 8 insectos durante 5 minutos, la distribución de cada aterrizaje se distribuye, independientemente de los demás, como una variable uniforme sobre el intervalo $[0, 5]$. En consecuencia, la probabilidad de que cada insecto hubiese aterrizado durante el primer minuto es $1/5$. Por lo tanto, la probabilidad de que exactamente 3 insectos hayan aterrizado durante el primer minuto es

$$\binom{8}{3} \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^5 = 56 \frac{4^5}{5^8} = 0.1468\dots$$

\square

1.5. Coloración y adelgazamiento de procesos de Poisson

Teorema 1.9 (Coloración). *Sea Π un proceso de Poisson de intensidad λ sobre \mathbb{R}^+ . Coloreamos los puntos de Π de la siguiente manera. Cada punto de Π se pinta de rojo con probabilidad p o de negro con probabilidad $1 - p$. Los puntos se pintan independientemente unos de otros. Sean Π_1 y Π_2 los conjuntos de puntos pintados de rojo y de negro, respectivamente. Entonces Π_1 y Π_2 son procesos de Poisson independientes de intensidades $p\lambda$ y $(1 - p)\lambda$, respectivamente.*

Demostración. Sea $t > 0$ fijo. Por la propiedad condicional, si $N(t) = n$, esos puntos tienen la misma distribución que n puntos independientes elegidos al azar sobre el intervalo $[0, t]$ de acuerdo con la distribución uniforme. Por tanto, podemos considerar n puntos elegidos al azar de esa manera. Por la independencia de los puntos, sus colores son independientes unos de los otros. Como la probabilidad de que un punto dado sea pintado de rojo es p y la probabilidad de sea pintado de negro es $1 - p$ se deduce que, condicional a $N(t) = n$, las cantidades $N_1(t)$ y $N_2(t)$ de puntos rojos y negros en $[0, t]$ tienen, conjuntamente, la distribución binomial

$$\mathbb{P}(N_1(t) = n_1, N_2(t) = n_2 | N(t) = n) = \frac{n!}{n_1!n_2!} p^{n_1} (1-p)^{n_2}, \text{ donde } n_1 + n_2 = n.$$

Por lo tanto, la probabilidad incondicional es

$$\begin{aligned} \mathbb{P}(N_1(t) = n_1, N_2(t) = n_2) &= \left(\frac{(n_1 + n_2)!}{n_1!n_2!} p^{n_1} (1-p)^{n_2} \right) \left(e^{-\lambda t} \frac{(\lambda t)^{n_1+n_2}}{(n_1 + n_2)!} \right) \\ &= \left(e^{-p\lambda t} \frac{(p\lambda t)^{n_1}}{n_1!} \right) \left(\frac{e^{-(1-p)\lambda t} ((1-p)\lambda t)^{n_2}}{n_2!} \right). \end{aligned}$$

Vale decir, las cantidades $N_1(t)$ y $N_2(t)$ de puntos rojos y negros en el intervalo $[0, t]$ son independientes y tienen distribuciones Poisson de intensidades $p\lambda t$ y $(1-p)\lambda t$, respectivamente.

La independencia de las contadoras de puntos en intervalos disjuntas sigue trivialmente del hecho de que Π tiene esa propiedad. \square

Otra prueba. Sean $N_1(t)$ y $N_2(t)$ la cantidad de arribos de tipo I y de tipo II que ocurren en $[0, t]$, respectivamente. Es claro que $N(t) = N_1(t) + N_2(t)$.

Los arribos de tipo I (II) son un proceso puntual aleatorio debido a que son una subsucesión (aleatoria) infinita de los arribos del proceso original y heredan su propiedad de independencia para intervalos disjuntos.

La prueba de que $\{N_1(t), t \geq 0\}$ y que $\{N_2(t), t \geq 0\}$ son procesos de Poisson independientes de intensidades $p\lambda$ y $(1-p)\lambda$, respectivamente, se completa observando que

$$\mathbb{P}(N_1(t) = n, N_2(t) = m) = \mathbb{P}(N_1(t) = n)\mathbb{P}(N_2(t) = m).$$

Condicionando a los valores de $N(t)$ y usando probabilidades totales se obtiene

$$\mathbb{P}(N_1(t) = n, N_2(t) = m) = \sum_{i=0}^{\infty} \mathbb{P}(N_1(t) = n, N_2(t) = m | N(t) = i)\mathbb{P}(N(t) = i)$$

Puesto que $\mathbb{P}(N_1(t) = n, N_2(t) = m | N(t) = i) = 0$ cuando $i \neq n + m$, la ecuación anterior se reduce a

$$\begin{aligned} \mathbb{P}(N_1(t) = n, N_2(t) = m) &= \mathbb{P}(N_1(t) = n, N_2(t) = m | N(t) = n + m)\mathbb{P}(N(t) = n + m) \\ &= \mathbb{P}(N_1(t) = n, N_2(t) = m | N(t) = n + m)e^{-\lambda t} \frac{(\lambda t)^{n+m}}{(n+m)!}. \end{aligned}$$

Dado que ocurrieron $n+m$ arribos, la probabilidad de que n sean de tipo I (y m sean de tipo

II) es la probabilidad binomial de que ocurran n éxitos en $n + m$ ensayos. Por lo tanto,

$$\begin{aligned}\mathbb{P}(N_1(t) = n, N_2(t) = m) &= \binom{n+m}{n} p^n (1-p)^m e^{-\lambda t} \frac{(\lambda t)^{n+m}}{(n+m)!} \\ &= \frac{(n+m)!}{n! m!} p^n (1-p)^m e^{-\lambda pt} e^{-\lambda(1-p)t} \frac{(\lambda t)^n (\lambda t)^m}{(n+m)!} \\ &= \left(e^{-\lambda pt} \frac{(\lambda pt)^n}{n!} \right) \left(e^{-\lambda(1-p)t} \frac{(\lambda(1-p)t)^m}{m!} \right).\end{aligned}$$

Lo que completa la demostración. \square

Ejemplo 1.10 (Insectos en un asado). Todo tipo de insectos aterrizan en la mesa de un asado a la manera de un proceso de Poisson de tasa 3 por minuto y cada insecto puede ser una mosca con probabilidad $2/3$, independientemente de la naturaleza de los demás insectos. Si a las 13:30 se sirven los chorizos, cuál es la probabilidad de que la tercer mosca tarde más de 2 minutos en aterrizar en la mesa?

Solución: Las moscas aterrizan en la mesa a la manera de un proceso de Poisson de tasa $\frac{2}{3} \cdot 3 = 2$ por minuto. En consecuencia, los aterrizajes de moscas ocurren cada tiempos exponenciales independientes de intensidad 2. De aquí se deduce que el tiempo que tarda en aterrizar la tercer mosca, S_3 tiene distribución $\Gamma(3, 2)$. Por lo tanto, la probabilidad de que la tercer mosca tarde más de 2 minutos en aterrizar en la mesa es

$$\mathbb{P}(S_3 > 2) = e^{-2 \cdot 2} \sum_{i=0}^{3-1} \frac{(2 \cdot 2)^i}{i!} = e^{-4} (1 + 4 + 8) = 0.2381\dots$$

Ejercicios adicionales

5. A un banco llegan clientes de acuerdo con un proceso de Poisson de intensidad 20 por hora. En forma independiente de los demás, cada cliente realiza un depósito con probabilidad $1/4$ o una extracción con probabilidad $3/4$.

(a) Si el banco abre sus puertas a las 10:00, cuál es la probabilidad de que el segundo depósito se efectué pasadas las 10:30?

(b) Cada depósito (en pesos) se distribuye como una variable $\mathcal{U}[100, 900]$ y cada extracción como una variable $\mathcal{U}[100, 500]$. Si un cliente realiza una operación bancaria de 200 pesos, cuál es la probabilidad de que se trate de un depósito?

1.6. Superposición de Procesos de Poisson: competencia

El siguiente teorema de superposición puede verse como complementario del teorema de coloración.

Teorema 1.11 (Superposición). *Sean Π_1 y Π_2 dos procesos de Poisson independientes de intensidades λ_1 y λ_2 , respectivamente, sobre \mathbb{R}^+ . El conjunto $\Pi = \Pi_1 \cup \Pi_2$ es un proceso de Poisson de intensidad $\lambda_1 + \lambda_2$.* \square

Demostración. Sean $N_1(t) = |\Pi_1 \cap [0, t]|$ y $N_2(t) = |\Pi_2 \cap [0, t]|$. Entonces $N_1(t)$ y $N_2(t)$ son variables aleatorias independientes con distribución Poisson de parámetros $\lambda_1 t$ y $\lambda_2 t$. Se infiere que la suma $N(t) = N_1(t) + N_2(t)$ tiene la distribución de Poisson de parámetro $\lambda_1 t + \lambda_2 t = (\lambda_1 + \lambda_2)t$. Más aún, si A_1, A_2, \dots , son intervalos disjuntos las variables aleatorias $N(A_1), N(A_2), \dots$ son independientes. Falta mostrar que, casi seguramente, $N(t) = |\Pi \cap [0, t]|$ para todo $t > 0$, que es lo mismo que decir que Π_1 y Π_2 no tienen puntos en común. Este es un paso técnico (ver el Lema 1.12) y la prueba puede omitirse en una primera lectura.

Lema 1.12. *Dos procesos de Poisson $\Pi_1 = \{S_n^1 : n \geq 0\}$ y $\Pi_2 = \{S_n^2 : n \geq 0\}$ independientes y de tasas λ_1 y λ_2 , respectivamente, no tienen puntos en común.*

Demostración. Basta probar que $\mathbb{P}(D(t)) = 0$ para todo t , donde $D(t)$ es el evento definido por

$$D(t) := \{\text{existen puntos en común en el intervalo } (0, t]\}$$

Para simplificar la notación lo demostrarímos para $D = D(1)$.

Sean $\{N_1(t), t \geq 0\}$ y $\{N_2(t), t \geq 0\}$ los procesos de conteo de los procesos de Poisson $\{S_n^1 : n \geq 0\}$ y $\{S_n^2 : n \geq 0\}$. El evento

$$D_n := \left\{ N_1 \left(\frac{i}{2^n}, \frac{i+1}{2^n} \right] + N_2 \left(\frac{i}{2^n}, \frac{i+1}{2^n} \right] \geq 2 \text{ para algún } i \in [0, 2^n - 1] \right\}$$

decrece a D cuando n tiende a infinito, y por lo tanto, por la continuidad de la probabilidad para sucesiones monótonas de eventos,

$$\mathbb{P}(D) = \lim_{n \rightarrow \infty} \mathbb{P}(D_n) = 1 - \lim_{n \rightarrow \infty} \mathbb{P}(D_n^c).$$

Pero

$$\begin{aligned} \mathbb{P}(D_n^c) &= \mathbb{P} \left(\bigcap_{i=1}^{2^n-1} \left\{ N_1 \left(\frac{i}{2^n}, \frac{i+1}{2^n} \right] + N_2 \left(\frac{i}{2^n}, \frac{i+1}{2^n} \right] \leq 1 \right\} \right) \\ &= \prod_{i=1}^{2^n-1} \mathbb{P} \left(N_1 \left(\frac{i}{2^n}, \frac{i+1}{2^n} \right] + N_2 \left(\frac{i}{2^n}, \frac{i+1}{2^n} \right] \leq 1 \right). \end{aligned}$$

Debido a que los procesos son temporalmente homogéneos, para cada i vale que

$$\mathbb{P} \left(N_1 \left(\frac{i}{2^n}, \frac{i+1}{2^n} \right] + N_2 \left(\frac{i}{2^n}, \frac{i+1}{2^n} \right] \leq 1 \right) = \mathbb{P} (N_1(2^{-n}) + N_2(2^{-n}) \leq 1)$$

Y el problema se reduce a calcular $\mathbb{P}(N_1(2^{-n}) + N_2(2^{-n}) \leq 1)$. La última probabilidad puede expresarse como la suma de los siguientes términos

$$\begin{aligned} \mathbb{P}(N_1(2^{-n}) = 0, N_2(2^{-n}) = 0) &= e^{-\lambda_1 2^{-n}} e^{-\lambda_2 2^{-n}}, \\ \mathbb{P}(N_1(2^{-n}) = 0, N_2(2^{-n}) = 1) &= e^{-\lambda_1 2^{-n}} e^{-\lambda_2 2^{-n}} \lambda_2 2^{-n}, \\ \mathbb{P}(N_1(2^{-n}) = 1, N_2(2^{-n}) = 0) &= e^{-\lambda_1 2^{-n}} \lambda_1 2^{-n} e^{-\lambda_2 2^{-n}}. \end{aligned}$$

En consecuencia,

$$\mathbb{P}(N_1(2^{-n}) + N_2(2^{-n}) \leq 1) = e^{-(\lambda_1 + \lambda_2)2^{-n}} (1 + (\lambda_1 + \lambda_2)2^{-n}). \quad (22)$$

Por lo tanto,

$$\mathbb{P}(D_n^c) = e^{-(\lambda_1 + \lambda_2)} (1 + (\lambda_1 + \lambda_2)2^{-n})^{2^n}. \quad (23)$$

La última cantidad tiende a 1 cuando $n \rightarrow \infty$, y se concluye que $\mathbb{P}(D) = 0$. \square

Teorema 1.13 (Competencia). *En la situación del Teorema 1.11, sea T el primer arribo del proceso $N = N_1 + N_2$ y J el índice del proceso de Poisson responsable por dicho arribo; en particular T es el primer arribo de N_J . Entonces*

$$\mathbb{P}(J = j, T \geq t) = \mathbb{P}(J = j)\mathbb{P}(T \geq t) = \frac{\lambda_j}{\lambda_1 + \lambda_2} e^{-(\lambda_1 + \lambda_2)t}.$$

En particular, J y T son independientes, $\mathbb{P}(J = j) = \frac{\lambda_j}{\lambda_1 + \lambda_2}$ y T tiene distribución exponencial de intensidad $\lambda_1 + \lambda_2$.

Demostración. Ver la demostración del Teorema que caracteriza la distribución del mínimo de dos exponenciales independientes. \square

Ejemplo 1.14 (Insectos en un asado). Moscas y abejas aterrizan en la mesa de un asado a la manera de dos procesos de Poisson independientes de tasas 2 y 1 por minuto, respectivamente. Cuál es la probabilidad de que el primer insecto en aterrizar en la mesa sea una mosca? Rta. 2/3. \square

1.7. Procesos de Poisson compuestos

Un proceso estocástico se dice un proceso de Poisson compuesto si puede representarse como

$$X(t) = \sum_{i=1}^{N(t)} Y_i$$

donde $\{N(t), t \geq 0\}$ es un proceso de Poisson, y las variables $\{Y_i, i \geq 1\}$ son iid e independientes de N .

Lema 1.15. Sea $X(t)$ un proceso de Poisson compuesto. Si $\{N(t), t \geq 0\}$ tiene intensidad λ y las variables Y tienen esperanza finita, entonces

$$\mathbb{E}[X(t)] = \lambda t \mathbb{E}[Y_1].$$

Más aún, si las variables Y tienen varianza finita, entonces,

$$\mathbb{V}(X(t)) = \lambda t \mathbb{E}[Y_1^2].$$

Demostración. Para calcular la esperanza de $X(t)$ condicionamos sobre $N(t)$:

$$\mathbb{E}[X(t)] = \mathbb{E}[\mathbb{E}[X(t) | N(t)]]$$

Ahora bien,

$$\begin{aligned}
\mathbb{E}[X(t) | N(t) = n] &= \mathbb{E}\left[\sum_{i=1}^{N(t)} Y_i | N(t) = n\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n Y_i | N(t) = n\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n Y_i\right] \quad \text{por la independencia de } Y_i \text{ y } N(t) \\
&= n\mathbb{E}[Y_1].
\end{aligned}$$

Esto implica que

$$\mathbb{E}[X(t) | N(t)] = N(t)\mathbb{E}[Y_1]$$

y por lo tanto,

$$\mathbb{E}[X(t)] = \mathbb{E}[N(t)\mathbb{E}[Y_1]] = \mathbb{E}[N(t)]\mathbb{E}[Y_1] = \lambda t\mathbb{E}[Y_1].$$

Aunque podemos obtener $E[X(t)^2]$ condicionando sobre $N(t)$, usaremos la fórmula de la varianza condicional

$$\mathbb{V}(X(t)) = \mathbb{E}[\mathbb{V}(X(t)|N(t))] + \mathbb{V}(\mathbb{E}[X(t)|N(t)]).$$

Ahora bien,

$$\begin{aligned}
\mathbb{V}[X(t) | N(t) = n] &= \mathbb{V}\left(\sum_{i=1}^{N(t)} Y_i | N(t) = n\right) \\
&= \mathbb{V}\left(\sum_{i=1}^n Y_i | N(t) = n\right) \\
&= \mathbb{V}\left(\sum_{i=1}^n Y_i\right) \quad \text{por la independencia de } Y_i \text{ y } N(t) \\
&= n\mathbb{V}[Y_1].
\end{aligned}$$

Esto implica que

$$\mathbb{V}(X(t) | N(t)) = N(t)\mathbb{V}(Y_1)$$

y por lo tanto,

$$\begin{aligned}
\mathbb{V}(X(t)) &= \mathbb{E}[N(t)\mathbb{V}(Y_1)] + \mathbb{V}(N(t)\mathbb{E}[Y_1]) \\
&= \mathbb{V}(Y_1)\mathbb{E}[N(t)] + \mathbb{E}[Y_1]^2\mathbb{V}(N(t)) \\
&= \mathbb{V}(Y_1)\lambda t + \mathbb{E}[Y_1]^2\lambda t \\
&= \lambda t\mathbb{E}[Y_1^2].
\end{aligned}$$

□

Ejemplo 1.16. Supongamos que la cantidad de accidentes en una fábrica industrial se rige por un proceso de Poisson de intensidad 4 por mes y que la cantidad de trabajadores damnificados en cada accidente son variables aleatorias independientes con distribución uniforme sobre $\{1, 2, 3\}$. Supongamos también que la cantidad de trabajadores damnificados en cada accidente es independiente de la cantidad de accidentes ocurridos. Se quiere hallar la media y la varianza de la cantidad anual de trabajadores damnificados en dicha fábrica.

Solución: Sean $N(t)$ la cantidad de accidentes en t meses e Y_i el número de trabajadores damnificados en el i -ésimo accidente, $i = 1, 2, \dots$. El número total de trabajadores damnificados en un año puede expresarse en la forma $X(12) = \sum_{i=1}^{N(12)} Y_i$.

Utilizando los resultados del Lema 1.15 tenemos que

$$\begin{aligned}\mathbb{E}[X(12)] &= (4 \cdot 12)\mathbb{E}[Y_1] = 48\mathbb{E}[Y_1] = 48 \cdot 2 = 96 \\ \mathbb{V}(X(12)) &= (4 \cdot 12)\mathbb{E}[Y_1^2] = 48 \cdot \frac{14}{3} = 224.\end{aligned}$$

□

Ejercicios adicionales

6. Una partícula suspendida en agua es bombardeada por moléculas en movimiento térmico de acuerdo con un proceso de Poisson de intensidad 10 impactos por segundo. Cuando recibe un impacto la partícula se mueve un milímetro hacia la derecha con probabilidad $3/4$ o un milímetro hacia la izquierda con probabilidad $1/4$. Transcurrido un minuto, cuál es la posición media de la partícula?
7. Un servidor recibe clientes de acuerdo con un proceso de Poisson de intensidad 4 clientes por hora. El tiempo de trabajo (en minutos) consumido en cada servicio es una variable aleatoria $\mathcal{U}[1, 9]$. Al cabo de 8 horas, cuál es el tiempo medio de trabajo consumido por todos los servicios?

2. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Brémaud, P.: *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, New York. (1999)
2. Feller, W.: *An introduction to Probability Theory and Its Applications*. Vol. 2. John Wiley & Sons, New York. (1971)
3. Ferrari, P. A., Galves, A.: *Construction of Stochastic Processes, Coupling and Regeneration*. (2001)
4. Grimmett, G. R., Stirzaker, D. R.: *Probability and Random Processes*. Oxford University Press, New York. (2001)

5. Kingman, J. F. K.: Poisson Processes. Oxford University Press. New York. (2002)
6. Meester, R.: A Natural Introduction to Probability Theory. Birkhauser, Berlin. (2008)
7. Ross, S.: Introduction to Probability Models. Academic Press, San Diego. (2007)

Normalidad y Teorema central del límite (Borradores, Curso 23)

Sebastian Grynberg

24 de abril de 2013



*¿dónde es más útil aplicar la fuerza de la propia voluntad:
en el desarrollo de la cantidad o en el de la calidad?
¿Cuál de los dos aspectos es más fiscalizable?
¿Cuál más fácilmente mensurable?
¿Sobre cuál se pueden hacer previsiones, construir planes de trabajo?*

Índice

1. La distribución normal	2
1.1. Presentación	2
1.2. Cuentas con normales	5
1.3. Ejemplos	6
1.4. Suma de normales independientes	7
2. Génesis de la distribución normal	8
2.1. Teorema límite de De Moivre - Laplace	8
3. Teorema central del límite	14
3.1. Ejemplos	15
4. Distribuciones relacionadas con la Normal	19
4.1. χ^2 (chi-cuadrado)	19
4.2. t de Student	21
4.3. F de Fisher	21
5. Bibliografía consultada	23

1. La distribución normal

1.1. Presentación

Definición 1.1. *La función definida por*

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad (1)$$

se llama la función densidad normal; su integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \quad (2)$$

es la función distribución normal.

Folclore. Se sabe que la función e^{-x^2} no admite una primitiva que pueda expresarse mediante un número finito de funciones elementales: x^ν , $\sin(x)$, $\cos(x)$, a^x , etc.... (Ver Piskunov, N., (1983). *cálculo diferencial e integral*, tomo I, Mir, Moscú). Sin embargo, usando técnicas de cambio de variables bidimensionales se puede demostrar que $\int_{-\infty}^{\infty} \varphi(x) dx = 1$.

La función $\Phi(x)$ crece desde 0 hasta 1. Su gráfico es una curva con forma de S con

$$\Phi(-x) = 1 - \Phi(x). \quad (3)$$

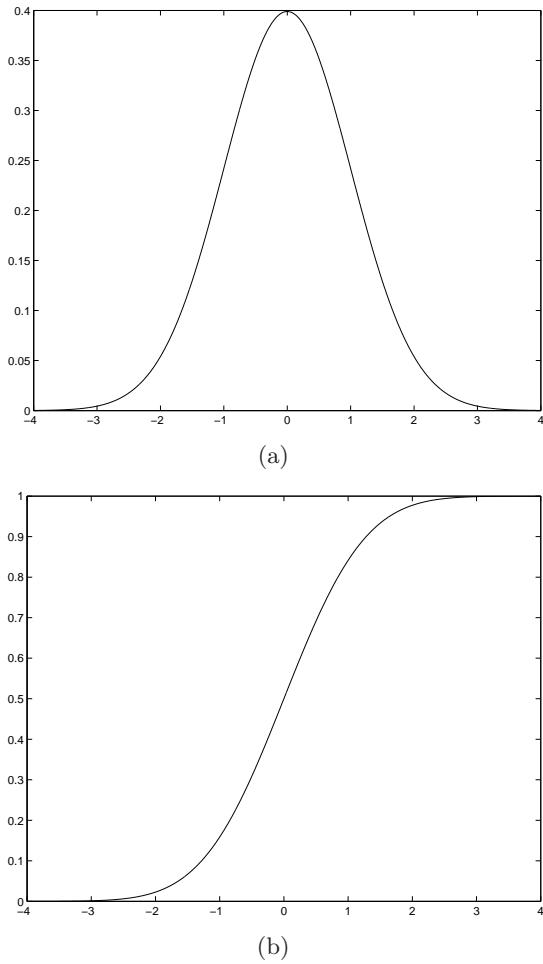


Figura 1: (a) La función densidad normal $\varphi(x) := \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$; (b) La función distribución normal $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$

Tablas. La tabla de valores de la función distribución normal se puede consultar en la mayoría de los libros sobre probabilidad y/o estadística. En general se tabulan los valores de $\Phi(x)$ para $x = d_0 + \frac{d_1}{10} + \frac{d_2}{100}$, donde $d_0 \in \{0, 1, 2, 3\}$ y $d_1, d_2 \in \{0, 1, 2, \dots, 9\}$. Las filas de la tabla están indexadas por los números $d_0.d_1$ y sus columnas por los números $0.0d_2$: en la posición $(d_0.d_1, 0.0d_2)$ de la tabla se encuentra el valor $\Phi(d_0.d_1 d_2)$. Por ejemplo, si se consulta la tabla del libro de Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, en fila 1.2 y columna de 0.08 puede leerse 0.8997, lo que significa que $\Phi(1.28) = 0.8997$.

En el Cuadro 1.1 reproducimos algunos de los valores de la tabla del Feller:

Lema 1.2. *Para cada $x > 0$ valen las siguientes desigualdades:*

$$\varphi(x) \left(\frac{1}{x} - \frac{1}{x^3} \right) < 1 - \Phi(x) < \varphi(x) \left(\frac{1}{x} \right). \quad (4)$$

x	1.28	1.64	1.96	2.33	2.58	3.09	3.29
$\Phi(x)$	0.8997	0.9495	0.975	0.9901	0.9951	0.9990	0.9995

Cuadro 1: En la tabla se muestran algunos valores de $\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$.

Demuestra. Usando que $\frac{d}{dx} \varphi(x) = -x\varphi(x)$ es fácil ver que las derivadas de los miembros de las desigualdades (4) satisfacen:

$$\begin{aligned} \frac{d}{dx} \left[\varphi(x) \left(\frac{1}{x} - \frac{1}{x^3} \right) \right] &= -\varphi(x) \left(1 - \frac{3}{x^4} \right). \\ \frac{d}{dx} [1 - \Phi(x)] &= -\varphi(x). \\ \frac{d}{dx} \left[\varphi(x) \left(\frac{1}{x} \right) \right] &= -\varphi(x) \left(1 + \frac{1}{x^2} \right). \end{aligned}$$

Por lo tanto,

$$\frac{d}{dx} \left[-\varphi(x) \left(\frac{1}{x} - \frac{1}{x^3} \right) \right] < \frac{d}{dx} [\Phi(x) - 1] < \frac{d}{dx} \left[-\varphi(x) \left(\frac{1}{x} \right) \right] \quad (5)$$

Las desigualdades (4) se obtienen integrando desde x hasta ∞ . \square

Nota Bene. De las desigualdades (4) se infiere un método de cálculo para aproximar los valores de $1 - \Phi(x)$: promediando los valores de los extremos de las desigualdades se obtiene una aproximación cuyo error absoluto es menor que la semi-diferencia entre ambos:

$$\left| 1 - \Phi(x) - \varphi(x) \left(\frac{1}{x} - \frac{1}{2x^3} \right) \right| \leq \frac{\varphi(x)}{2x^3}. \quad (6)$$

De la desigualdad (6) se puede ver que la aproximación

$$\Phi(x) \approx 1 - \varphi(x) \left(\frac{1}{x} - \frac{1}{2x^3} \right) \quad (7)$$

es prácticamente inútil para valores “pequeños” de x (i.e., $x \in (0, 1]$) pero va mejorando a medida que los valores de x “crecen”. Usando la aproximación dada en (7) se obtienen las siguientes aproximaciones

x	1.28	1.64	1.96	2.33	2.58	3.09	3.29
$\Phi(x)$	0.90454	0.94839	0.97406	0.98970	0.99487	0.99896	0.99948
$ \text{error} \leq$	0.04192	0.01178	0.00388	0.00104	0.00041	0.00005	0.00002

Cuadro 2: Algunos valores de $\Phi(x)$ obtenidos mediante la estimación (7).

Nota histórica La distribución normal fue descubierta por De Moivre en 1733 como resultado de analizar la forma límite de la distribución binomial simétrica y redescubierta nuevamente por Gauss (1809) y Laplace (1812) quienes la estudiaron en relación con sus trabajos sobre la teoría de los errores de observación. Laplace dio, además, el primer enunciado (incompleto) del teorema central del límite. (Ver Cramer, H., (1970). *Métodos matemáticos de estadística*, Aguilar, Madrid.)

1.2. Cuentas con normales

Sean $\mu \in \mathbb{R}$ y $\sigma > 0$ arbitrarios, pero fijos. Se dice que la variable aleatoria X tiene distribución normal de parámetros μ y σ^2 y se denota $X \sim \mathcal{N}(\mu, \sigma^2)$ si la función densidad de X es de la forma

$$\varphi_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (8)$$

Nota Bene. Un hecho importante sobre las variables aleatorias normales es que si X tiene distribución normal $\mathcal{N}(\mu, \sigma^2)$, entonces

$$Z = \frac{X - \mu}{\sigma} \quad (9)$$

tiene distribución normal $\mathcal{N}(0, 1)$. En efecto,

$$\begin{aligned} \mathbb{P}(Z \leq z) &= \mathbb{P}((X - \mu)/\sigma \leq z) = \mathbb{P}(X \leq z\sigma + \mu) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{z\sigma+\mu} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{1}{2}t^2} dt \quad \text{por sustitución } x = t\sigma + \mu. \end{aligned}$$

Este hecho significa que si trasladamos el origen de las abscisas en μ y cambiamos la escala de manera tal que σ represente la unidad de medida, la distribución normal $\mathcal{N}(\mu, \sigma^2)$ se transforma en la distribución normal $\mathcal{N}(0, 1)$. Su importancia práctica radica en que permite reducir el cálculo de probabilidades de las distribuciones normales $\mathcal{N}(\mu, \sigma^2)$ al de la distribución normal $\mathcal{N}(0, 1)$. Motivo por el cual esta última recibe el nombre de *normal estándar* (o *típica*). Más precisamente, si X tiene distribución normal $\mathcal{N}(\mu, \sigma^2)$, su función de distribución podrá reducirse a la función de distribución normal $\Phi(\cdot)$ definida en (2) de la siguiente manera:

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right). \quad (10)$$

La identidad (10) resume toda la información probabilísticamente relevante sobre la variable aleatoria $X \sim \mathcal{N}(\mu, \sigma^2)$ y permite calcular (con ayuda de la tabla de la función de distribución normal $\Phi(\cdot)$) la probabilidad de que la variable X se encuentre en cualquier intervalo prefijado de antemano:

$$\mathbb{P}(a < X < b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right). \quad (11)$$

En particular, cuando el intervalo (a, b) es simétrico con respecto a μ , las cantidades a y b se pueden expresar en la forma $a = \mu - \epsilon$, $b = \mu + \epsilon$, donde $\epsilon > 0$, y la fórmula (11) adopta la forma

$$\mathbb{P}(|X - \mu| < \epsilon) = \Phi\left(\frac{\epsilon}{\sigma}\right) - \Phi\left(-\frac{\epsilon}{\sigma}\right) = 2\Phi\left(\frac{\epsilon}{\sigma}\right) - 1. \quad (12)$$

Significado de los parámetros μ y σ^2 . La relación (9) dice que si X es una variable aleatoria con distribución normal de parámetros μ y σ^2 , entonces $X = \sigma Z + \mu$ donde Z es una variable con distribución normal estándar. Cálculos de rutina muestran que $\mathbb{E}[Z] = 0$ y $\mathbb{V}(Z) = 1$, lo que permite deducir que *la media y la varianza de la $\mathcal{N}(\mu, \sigma^2)$ son μ y σ^2 , respectivamente.*

1.3. Ejemplos

Ejemplo 1.3. Una maquina produce ejes cuyos diámetros X tienen distribución normal de media $\mu = 10$ mm y varianza $\sigma^2 = 0.25$ mm. Un eje se considera defectuoso si $X < 9.5$ mm. Cuál es la probabilidad de que un eje elegido al azar resulte defectuoso?

Solución: El problema se resuelve calculando $\mathbb{P}(X < 9.5)$. Poniendo $\mu = 10$ y $\sigma = 0.5$ en la fórmula (10) obtenemos $\mathbb{P}(X < 9.5) = \Phi\left(\frac{9.5-10}{0.5}\right) = \Phi(-1) = 0.1587$. \square

Curva peligrosa. De inmediato podría surgir una objeción al uso de la distribución normal $\mathcal{N}(10, 0.25)$ para modelar el diámetro de los ejes. Al fin y al cabo, los diámetros deben ser positivos y la distribución normal adopta valores positivos y negativos. Sin embargo, el modelo anterior asigna una probabilidad despreciable al evento $X < 0$. En efecto, $\mathbb{P}(X < 0) = \mathbb{P}\left(\frac{X-10}{0.5} < \frac{0-10}{0.5}\right) = \mathbb{P}(Z < -20) = \Phi(-20) = 1 - \Phi(20)$. De acuerdo con la estimación (6) tenemos que $1 - \Phi(20) \approx \varphi(20)\left(\frac{1}{20} - \frac{1}{2 \cdot 20^3}\right) = O(10^{-89})$. Este tipo de situación es habitual en la práctica. Se tiene una variable aleatoria X de la que se sabe que no puede tomar valores negativos (p.ej. una distancia, una longitud, un área, un peso, una temperatura, un precio, etc.) y se la modela utilizando una distribución normal $\mathcal{N}(\mu, \sigma^2)$; motivados, por ejemplo, por cuestiones de simetría. En principio, el modelo podrá ser perfectamente válido siempre y cuando los valores de los parámetros μ y σ^2 sean tales que la probabilidad $\mathbb{P}(X < 0)$ sea prácticamente 0.

Nota Bene sobre grandes desvíos. Sea X una variable aleatoria con distribución normal de media μ y varianza σ^2 . Sea $t > 0$, utilizando la fórmula (12) podemos ver que

$$p_t := \mathbb{P}(|X - \mu| > t\sigma) = 1 - \mathbb{P}(|X - \mu| \leq t\sigma) = 1 - \left(2\Phi\left(\frac{t\sigma}{\sigma}\right) - 1\right) = 2(1 - \Phi(t)).$$

Usando la tabla de la distribución normal $\Phi(\cdot)$ se puede ver que $p_1 = 0.3174$, $p_2 = 0.0454$, $p_3 = 0.0028$. Estos probabilidades admiten la siguiente interpretación: cerca del 32% de los valores de una variable $X \sim \mathcal{N}(\mu, \sigma^2)$ se desvían de su media en más de σ ; solamente cerca de un 5% lo hacen en más de 2σ y solamente cerca de un 3% en más de 3σ . Esto da lugar a que en la mayor parte de los problemas de la práctica se consideren casi imposibles las desviaciones respecto de la media μ que superen 3σ y se consideren limitados por el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$ todos los valores prácticamente posibles de la variable X .

Ejemplo 1.4. Sea X una variable aleatoria con distribución normal de media $\mu = 3$ y varianza $\sigma^2 = 4$. ¿Cuál es la probabilidad de que X sea no menor que 1 y no mayor que 7?

Solución: Poner $\mu = 3$ y $\sigma = 2$ en la fórmula (11) y usar la tabla de la distribución normal $\Phi(\cdot)$: $\mathbb{P}(1 \leq X \leq 7) = \Phi\left(\frac{7-3}{2}\right) - \Phi\left(\frac{1-3}{2}\right) = \Phi(2) - \Phi(-1) = 0.9773 - 0.1587 = 0.8186$. \square

1.4. Suma de normales independientes

Lema 1.5. Sean X_1 y X_2 dos variables aleatorias independientes con distribución normal $\mathcal{N}(\mu_1, \sigma_1^2)$ y $\mathcal{N}(\mu_2, \sigma_2^2)$, respectivamente. Entonces $X_1 + X_2$ tiene distribución normal $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Demostración. Observando que $X_1 + X_2 = (X_1 - \mu_1) + (X_2 - \mu_2) + \mu_1 + \mu_2$ el problema se reduce a considerar el caso $\mu_1 = \mu_2 = 0$. La prueba se obtiene mostrando que la convolución de las densidades $f_1(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp(-x_1^2/2\sigma_1^2)$ y $f_2(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp(-x_2^2/2\sigma_2^2)$ es la densidad normal de media $\mu_1 + \mu_2$ y varianza $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Por definición

$$(f_1 * f_2)(x) = \int_{-\infty}^{\infty} f_1(x-y)f_2(y) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-y)^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}\right) dy \quad (13)$$

El resultado se obtendrá mediante un poco de álgebra, bastante paciencia, y un cambio de variables en la integral del lado derecho de la identidad (13).

$$\begin{aligned} \exp\left(-\frac{(x-y)^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}\right) &= \exp\left(-\frac{1}{2} \left(\frac{\sigma}{\sigma_1\sigma_2}y - \frac{\sigma_2}{\sigma\sigma_1}x \right)^2 - \frac{x^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2} \left(\frac{\sigma}{\sigma_1\sigma_2}y - \frac{\sigma_2}{\sigma\sigma_1}x \right)^2\right) \exp\left(-\frac{x^2}{2\sigma^2}\right) \end{aligned}$$

La primera igualdad se obtuvo completando cuadrados respecto de y en la expresión $-\frac{(x-y)^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}$ y reagrupando algunos términos. Mediante el cambio de variables $z = \frac{\sigma}{\sigma_1\sigma_2}y - \frac{\sigma_2}{\sigma\sigma_1}x$, cuya diferencial es de la forma $dz = \frac{\sigma}{\sigma_1\sigma_2}dy$, se puede ver que

$$(f_1 * f_2)(x) = \frac{1}{2\pi\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

□

Este resultado se puede generalizar para una suma de n variables aleatorias independientes: Sean X_1, X_2, \dots, X_n variables aleatorias independientes con distribuciones normales: $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $1 \leq i \leq n$. Entonces,

$$\sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

La prueba se obtiene por inducción y utilizando la siguiente propiedad “hereditaria” de familias de variables aleatorias independientes (cuya prueba puede verse en el Capítulo 1 del libro de Durrett, R.(1996): *Probability Theory and Examples*): Si X_1, X_2, \dots, X_n son variables aleatorias independientes, entonces funciones (medibles) de familias disjuntas de las X_i también son independientes.

Nota Bene. Observando que para cada $a \in \mathbb{R}$ y $X \sim \mathcal{N}(\mu, \sigma^2)$ resulta que $aX \sim \mathcal{N}(a\mu, a^2\sigma^2)$ se obtiene el siguiente resultado:

Teorema 1.6. Sean X_1, X_2, \dots, X_n variables aleatorias independientes con distribuciones normales: $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $1 \leq i \leq n$ y sean a_1, a_2, \dots, a_n números reales cualesquiera. Entonces,

$$\sum_{i=1}^n a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right).$$

2. Génesis de la distribución normal

2.1. Teorema límite de De Moivre - Laplace

En 1733, De Moivre observó que la distribución binomial correspondiente a la cantidad de éxitos, S_n , en n ensayos de Bernoulli simétricos tiene la forma límite de una campana. Esta observación fue la clave que le permitió descubrir la famosa *campana de Gauss* y allanar el camino que lo condujo a establecer la primera versión del *Teorema Central del Límite*: la convergencia de la distribución Binomial($n, 1/2$) a la distribución normal estándar. En 1801, Laplace refinó y generalizó este resultado al caso de la distribución Binomial(n, p). El Teorema de De Moivre-Laplace, que enunciamos más abajo, mejora sustancialmente la Ley débil de los grandes números porque proporciona una estimación mucho más precisa de las probabilidades $\mathbb{P}(|\frac{S_n}{n} - p| \leq \epsilon)$.

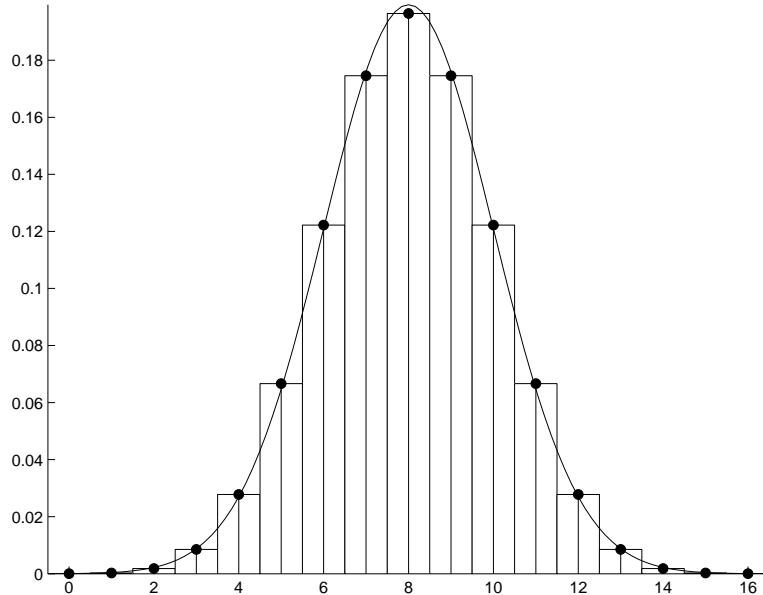


Figura 2: Relación entre la distribución Binomial simétrica y la distribución normal. La probabilidad de que ocurran k éxitos en n ensayos de Bernoulli está representada por un segmento paralelo al eje de las abscisas localizado en la ordenada k de altura igual a $\mathbb{P}(S_n = k)$. La curva continua “aproxima” los valores de $\mathbb{P}(S_n = k)$. Observar que dichas probabilidades también se pueden representar como áreas de rectángulos de altura $\mathbb{P}(S_n = k)$ y de base unitaria centrada en k .

Teorema 2.1 (Teorema límite de De Moivre-Laplace). *Consideramos una sucesión de ensayos de Bernoulli independientes. Sean p la probabilidad de éxito en cada ensayo y S_n la cantidad de éxitos observados en los primeros n ensayos. Para cualquier $x \in \mathbb{R}$ vale que*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq x \right) = \Phi(x), \quad (14)$$

donde $\Phi(x) := \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ es la función distribución normal estándar.

Demostración. Ver Capítulo VII de Feller, W., (1971). *An Introduction to Probability Theory and Its Applications*, Vol. I, John Wiley & Sons, New York. \square

¿Qué significa el Teorema Límite de De Moivre-Laplace? Para contestar esta pregunta vamos a reconstruir las ideas principales de su génesis. En otras palabras, vamos a (re)construir el Teorema. La clave de la construcción está “embutida” en la Figura 2. La imagen permite “capturar de inmediato” la existencia de una forma límite para la distribución Binomial en el caso simétrico $p = 1/2$.

Paso 1. El primer paso en la dirección del Teorema de De Moivre consiste en darse cuenta que la Figura 2 señala la existencia de una forma límite. En una primera fase (completamente abstracta) podemos conjeturar que “la distribución binomial simétrica tiene una forma asintótica. En otras palabras, cuando la cantidad de ensayos de Bernoulli es suficientemente grande, salvo traslaciones y cambios de escala apropiados, la distribución Binomial se parece a una función continua par, $\varphi(x)$, cuyo gráfico tiene la forma de una campana.”

Paso 2. El segundo paso consiste en precisar la naturaleza de la traslación y los cambios de escala que permiten “capturar” esa forma límite. Si se reflexiona sobre el significado de la media y la varianza de una variable aleatoria, parece claro que la forma límite se obtendrá centrando la variable S_n en su valor medio, $\mathbb{E}[S_n] = \frac{1}{2}n$, y adoptando como unidad de medida la desviación típica de los valores observados respecto de dicho valor, $\sigma(S_n) = \frac{1}{2}\sqrt{n}$. El significado geométrico de esta transformación consiste en (1) trasladar el origen de las abscisas en $\frac{1}{2}n$ y (2) dividirlas por $\frac{1}{2}\sqrt{n}$. Para que las áreas de los rectángulos sigan representando probabilidades, las ordenadas deben multiplicarse por el mismo número. Este paso permite enunciar la siguiente versión mejorada de la conjetura inicial: “existe una función continua $\varphi(x)$ tal que

$$\mathbb{P}(S_n = k) = \binom{n}{k} \left(\frac{1}{2}\right)^n \sim \frac{1}{\frac{1}{2}\sqrt{n}} \varphi\left(\frac{k - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}\right), \quad (15)$$

siempre y cuando n sea suficientemente grande.”

Paso 3. Establecida la conjetura el problema consiste en “descubrir” la expresión de la función $\varphi(x)$ y en precisar cuál es el sentido de la relación aproximada que aparece en (15). En este punto no queda otra que “arremangarse y meter la mano en el barro”. Como resultado se obtiene que la expresión de la función $\varphi(x)$ es

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

y que la relación \sim vale para valores de k del orden de \sqrt{n} y significa que el cociente de los dos lados tiende a 1 cuando $n \rightarrow \infty$.

Nota Bene. La relación (15) expresa matemáticamente un hecho que se observa claramente en la Figura 2: la campana “pasa” por los puntos de base k y altura $\mathbb{P}(S_n = k)$. Conviene observar que la expresión que aparece en el lado derecho de la relación (15) es la función de densidad de la normal $\mathcal{N}(\frac{1}{2}n, \frac{1}{4}n)$ evaluada en $x = k$. En la práctica, esto significa que para obtener una buena aproximación de la probabilidad de observar k éxitos en n ensayos de Bernoulli independientes, basta con evaluar la densidad de la normal $\mathcal{N}(\frac{1}{2}n, \frac{1}{4}n)$ en $x = k$. Sin temor a equivocarnos, podemos resumir estas observaciones mediante una expresión de la forma $S_n \sim \mathcal{N}(\mathbb{E}[S_n], \mathbb{V}(S_n))$.

Paso 4. Observar que para cada $x_1 < x_2$ vale que

$$\begin{aligned}\mathbb{P}\left(x_1 \leq \frac{S_n - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} \leq x_2\right) &= \mathbb{P}\left(\frac{1}{2}n + x_1 \frac{1}{2}\sqrt{n} \leq S_n \leq \frac{1}{2}n + x_2 \frac{1}{2}\sqrt{n}\right) \\ &= \sum_{x_1 \frac{1}{2}\sqrt{n} \leq j \leq x_2 \frac{1}{2}\sqrt{n}} \mathbb{P}\left(S_n = \frac{1}{2}n + j\right) \\ &\approx \sum_{x_1 \leq jh \leq x_2} h\varphi(jh),\end{aligned}\tag{16}$$

donde $h = \frac{2}{\sqrt{n}}$ y la suma se realiza sobre todos los enteros j tales que $x_1 \leq jh \leq x_2$. Cada uno de los sumandos que aparecen en el lado derecho de la aproximación (16) es el área de un rectángulo de base $[kh, (k+1)h]$ y altura $\varphi(kh)$. Como la función $\varphi(\cdot)$ es continua, para valores pequeños de h la suma total de las áreas de los rectángulos debe estar próxima al área bajo la curva de la densidad normal entre x_1 y x_2 . Por lo tanto, debe valer lo siguiente

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(x_1 \leq \frac{S_n - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} \leq x_2\right) = \int_{x_1}^{x_2} \varphi(t)dt = \Phi(x_2) - \Phi(x_1).\tag{17}$$

Este paso puede hacerse formalmente preciso “arremangándose y metiendo la mano en ...”

Nota Bene. La variable aleatoria que aparece dentro de la probabilidad del lado izquierdo de (17)

$$S_n^* = \frac{S_n - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} = \frac{S_n - \mathbb{E}[S_n]}{\sigma(S_n)}\tag{18}$$

es una medida de la desviación de S_n respecto de la media $\mathbb{E}[S_n]$ en unidades de la desviación típica $\sigma(S_n)$. El teorema límite de De Moivre-Laplace significa que cuando se considera una cantidad n (suficientemente grande) de ensayos de Bernoulli independientes, la distribución de la variable aleatoria S_n^* es “prácticamente indistinguible” de la distribución normal estándar $\mathcal{N}(0, 1)$.

Comentario sobre prueba del Teorema 2.1. Si se sigue con cuidado la demostración presentada por Feller se puede ver que las herramientas principales de la prueba son el desarrollo de Taylor (1712) de la función $\log(1+t) = t + O(t^2)$ y la fórmula asintótica de Stirling (1730) para los números factoriales $n! \sim \sqrt{2\pi n} n^n e^{-n}$. Partiendo de la función de probabilidad de la Binomial($n, 1/2$) se “deduce” la expresión de la función densidad normal $(\sqrt{2\pi})^{-1} e^{-x^2/2}$: el factor $(\sqrt{2\pi})^{-1}$ proviene de la fórmula de Stirling y el factor $e^{-x^2/2}$ del desarrollo de Taylor. Dejando de lado los recursos técnicos utilizados en la prueba, se observa que las ideas involucradas son simples y “recorren el camino del descubrimiento” de De Moivre (1733).

Ejemplo 2.2. Se lanza 40 veces una moneda honesta. Hallar la probabilidad de que se obtengan exactamente 20 caras. Usar la aproximación normal y compararla con la solución exacta.

Solución: La cantidad de caras en 40 lanzamientos de una moneda honesta, S_{40} , es una variable Binomial de parámetros $n = 40$ y $p = 1/2$. La aproximación normal (15) establece que

$$\mathbb{P}(S_{40} = 20) \sim \frac{1}{\frac{1}{2}\sqrt{40}} \varphi(0) = \frac{1}{\sqrt{20\pi}} = 0.12615\dots$$

El resultado exacto es

$$\mathbb{P}(X = 20) = \binom{40}{20} \left(\frac{1}{2}\right)^{40} = 0.12537\dots$$

□

Ejemplo 2.3. Se dice que los recién nacidos de madres fumadoras tienden a ser más pequeños y propensos a una variedad de dolencias. Se conjectura que además parecen deformes. A un grupo de enfermeras se les mostró una selección de fotografías de bebés, la mitad de los cuales nacieron de madres fumadoras; las enfermeras fueron invitadas a juzgar a partir de la apariencia de cada uno si la madre era fumadora o no. En 1500 ensayos se obtuvieron 910 respuestas correctas. La conjectura es plausible?

Solución: Aunque superficial, un argumento atendible consiste en afirmar que, si todos los bebés parecen iguales, la cantidad de repuestas correctas S_n en n ensayos es una variable aleatoria con distribución Binomial ($n, 1/2$). Entonces, para n grande

$$\mathbb{P}\left(\frac{S_n - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} > 3\right) = 1 - \mathbb{P}\left(\frac{S_n - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} \leq 3\right) \approx 1 - \Phi(3) \approx \frac{1}{1000}$$

por el Teorema límite de De Moivre-Laplace. Para los valores dados de S_n ,

$$\frac{S_n - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} = \frac{910 - 750}{5\sqrt{15}} \approx 8.$$

Se podría decir que el evento $\{X - \frac{1}{2}n > \frac{3}{2}\sqrt{n}\}$ es tan improbable que su ocurrencia arroja dudas sobre la suposición original de que los bebés parecen iguales. Este argumento otorgaría cierto grado de credibilidad a la conjectura enunciada. □

Comentarios sobre el caso general

1. En el caso general, la probabilidad de éxito en cada ensayo de Bernoulli individual es $p \in (0, 1)$. Si S_n es la cantidad de éxitos observados en los primeros n ensayos, entonces $\mathbb{E}[S_n] = np$ y $\mathbb{V}(S_n) = np(1 - p)$. Por lo tanto, la variable aleatoria

$$S_n^* := \frac{S_n - np}{\sqrt{np(1 - p)}} \tag{19}$$

es una medida de la desviación de S_n respecto de la media $\mathbb{E}[S_n] = np$ en unidades de la desviación típica $\sigma(S_n) = \sqrt{np(1 - p)}$. El teorema límite de De Moivre-Laplace significa

que cuando se considera una cantidad n (suficientemente grande) de ensayos de Bernoulli independientes, la distribución de la variable aleatoria S_n^* es “prácticamente indistinguible” de la distribución normal estándar $\mathcal{N}(0, 1)$.

2. Técnicamente la prueba del teorema se puede hacer recurriendo a las mismas herramientas utilizadas en la prueba del caso simétrico, pero los cálculos involucrados son más complicados. Sin embargo, el resultado también es claro si se observan las gráficas de la distribución Binomial(n, p). En la Figura 3 se ilustra el caso $n = 16$ y $p = 1/4$. Nuevamente es “evidente” que la forma límite de distribución Binomial debe ser la distribución normal.

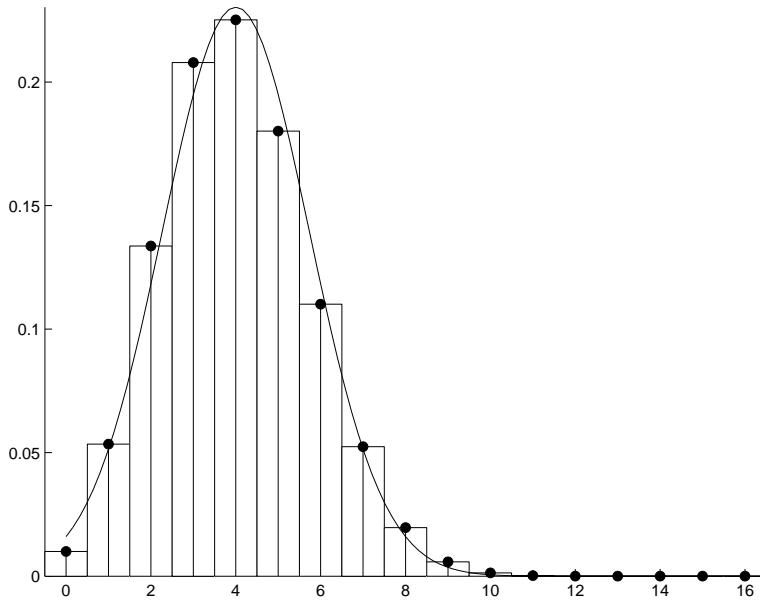


Figura 3: Gráfica de la función de probabilidad binomial con $n = 16$ y $p = 1/4$. Cerca del término central $m = np = 4$, salvo un cambio de escala (cuya unidad de medida es $\sqrt{np(1-p)} = \sqrt{3}$) la gráfica es “indistinguible” de la gráfica de la densidad normal.

3. De la Figura 3 debería estar claro que, para n suficientemente grande, debe valer lo siguiente

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{\sqrt{np(1-p)}} \varphi \left(\frac{k - np}{\sqrt{np(1-p)}} \right). \quad (20)$$

Ejemplo 2.4. Para el caso ilustrado en la Figura 3: $n = 16$ y $p = 1/4$, la aproximación (20) es bastante buena, incluso con un valor de n pequeño. Para $k = 0, \dots, 4$ las probabilidades $\mathbb{P}(S_n = 4+k)$ son 0.2252, 0.1802, 0.1101, 0.0524, 0.0197. Las aproximaciones correspondientes son 0.2303, 0.1950, 0.1183, 0.0514, 0.0160. \square

Nota Bene. El Teorema límite de De Moivre-Laplace justifica el uso de los métodos de la curva normal para aproximar probabilidades relacionadas con ensayos de Bernoulli con probabilidad de éxito p . La experiencia “indica” que la aproximación es bastante buena siempre que $np > 5$ cuando $p \leq 1/2$, y $n(1-p)$ cuando $p > 1/2$. Un valor muy pequeño de p junto con un valor de n moderado darán lugar a una media pequeña y con ello se obtendrá una

distribución asimétrica. La mayor parte de la distribución se acumulará alrededor de 0, impiendo con ello que una curva normal se le ajuste bien. Si la media se aparta por lo menos 5 unidades de una y otra extremidad, la distribución tiene suficiente espacio para que resulte bastante simétrica. (Ver la Figura 4).

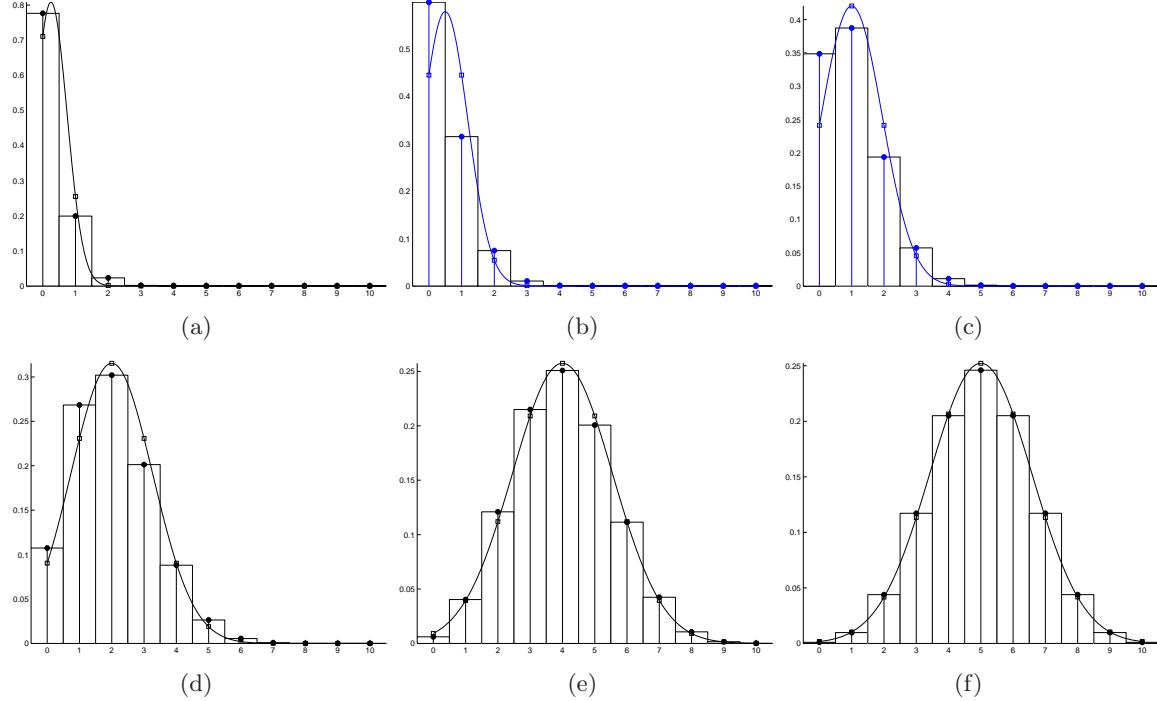


Figura 4: Comparación entre la distribución Binomial($10, p$) y su aproximación por la normal para distintos valores de p (a) $p = 0.025$; (b) $p = 0.05$; (c) $p = 0.1$; (d) $p = 0.2$; (e) $p = 0.4$; (f) $p = 0.5$.

Ejemplo 2.5 (Encuesta electoral). Queremos estimar la proporción del electorado que prefiere votar a un cierto candidato. Para ello consideramos que el voto de cada elector tiene una distribución Bernoulli de parámetro p . Concretamente, queremos encontrar un tamaño muestral n suficiente para que con una certeza del 99.99 % podamos garantizar un error máximo de 0.02 entre el verdadero valor de p y la proporción muestral S_n/n . En otras palabras, queremos encontrar n tal que

$$\mathbb{P} \left(\left| \frac{S_n}{n} - p \right| \leq 0.02 \right) \geq 0.9999. \quad (21)$$

Para acotar la incertezas usaremos la aproximación por la normal provista por el teorema límite de De Moivre - Laplace. Para ello, en lugar de observar la variable S_n , debemos observar la variable normalizada $S_n^* := (S_n - np)/\sqrt{np(1-p)}$. En primer lugar observamos que, como consecuencia del teorema límite, tenemos la siguiente aproximación

$$\mathbb{P} \left(\left| \frac{S_n - np}{\sqrt{np(1-p)}} \right| \leq a \right) \approx \Phi(-a) - \Phi(a) = 2\Phi(a) - 1 \quad (22)$$

o lo que es equivalente

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \leq \frac{a\sqrt{p(1-p)}}{\sqrt{n}}\right) \approx 2\Phi(a) - 1. \quad (23)$$

Como el verdadero valor de p es desconocido, la fórmula (23) no puede aplicarse directamente ya que no se conoce el valor de $\sqrt{p(1-p)}$. Sin embargo, es fácil ver que $\sqrt{p(1-p)} \leq 1/2$ y por lo tanto

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \leq \frac{a}{2\sqrt{n}}\right) \geq \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \leq \frac{a\sqrt{p(1-p)}}{\sqrt{n}}\right) \approx 2\Phi(a) - 1. \quad (24)$$

Esta última relación es la herramienta con la que podemos resolver nuestro problema.

En primer lugar tenemos que resolver la ecuación $2\Phi(a) - 1 = 0.9999$ o la ecuación equivalente $\Phi(a) = \frac{1.9999}{2} = 0.99995$. La solución de esta ecuación se obtiene consultando una tabla de la distribución normal: $a = 3.9$. Reemplazando este valor de a en (24) obtenemos

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \leq \frac{3.9}{2\sqrt{n}}\right) \geq 0.9999.$$

En segundo lugar tenemos que encontrar los valores de n que satisfacen la desigualdad

$$\frac{3.9}{2\sqrt{n}} \leq 0.02. \quad (25)$$

Es fácil ver que n satisface la desigualdad (25) si y solo si

$$n \geq \left(\frac{3.9}{0.04}\right)^2 = (97.5)^2 = 9506.2$$

El problema está resuelto. □

3. Teorema central del límite

Los teoremas sobre normalidad asintótica de sumas de variables aleatorias se llaman Teoremas Centrales del Límite. El Teorema límite de De Moivre - Laplace es un Teorema Central del Límite para variables aleatorias independientes con distribución Bernoulli(p). Una versión más general es la siguiente:

Teorema 3.1 (Teorema Central del Límite). *Sea X_1, X_2, \dots una sucesión de variables aleatorias independientes idénticamente distribuidas, cada una con media μ y varianza σ^2 . Entonces la distribución de*

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$$

tiende a la normal estándar cuando $n \rightarrow \infty$. Esto es,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x),$$

donde $\Phi(x) := \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ es la función de distribución de una normal de media 0 y varianza 1.

Demostración. Ver Capítulo XV de Feller, W., (1971). *An Introduction to Probability Theory and Its Applications*, Vol. II, John Wiley & Sons, New York. \square

Corolario 3.2. Sea X_1, X_2, \dots una sucesión de variables aleatorias independientes idénticamente distribuidas, cada una con media μ y varianza σ^2 . Si n es suficientemente grande, para cada valor $a > 0$ vale la siguiente aproximación

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \leq a \frac{\sigma}{\sqrt{n}}\right) \approx 2\Phi(a) - 1 \quad (26)$$

Demostración. El teorema central del límite establece que si n es suficientemente grande, entonces para cada $x \in \mathbb{R}$ vale que

$$\mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) \approx \Phi(x) \quad (27)$$

De la aproximación (27) se deduce que para cada valor $a > 0$

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}\right| \leq a\right) \approx \Phi(a) - \Phi(-a) = 2\Phi(a) - 1. \quad (28)$$

El resultado se obtiene de (28) observando que

$$\left|\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}\right| = \frac{n}{\sigma\sqrt{n}} \left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| = \frac{\sqrt{n}}{\sigma} \left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right|. \quad (29)$$

\square

Nota Bene. Para los usos prácticos, especialmente en estadística, el resultado límite en sí mismo no es de interés primordial. Lo que interesa es usarlo como una aproximación con valores finitos de n . Aunque no es posible dar un enunciado consiso sobre cuan buena es la aproximación, se pueden dar algunas pautas generales y examinando algunos casos especiales se puede tener alguna idea más precisa del comportamiento de cuan buena es la aproximación. Qué tan rápido la aproximación es buena depende de la distribución de los sumandos. Si la distribución es bastante simétrica y sus colas decaen rápidamente, la aproximación es buena para valores relativamente pequeños de n . Si la distribución es muy asimétrica o si sus colas decaen muy lentamente, se necesitan valores grandes de n para obtener una buena aproximación.

3.1. Ejemplos

Ejemplo 3.3 (Suma de uniformes). Puesto que la distribución uniforme sobre $[-\frac{1}{2}, \frac{1}{2}]$ tiene media 0 y varianza $\frac{1}{12}$, la suma de 12 variables independientes $\mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$ tiene media 0 y varianza 1. La distribución de esa suma está muy cerca de la normal.

Ejemplo 3.4. Para simplificar el cálculo de una suma se redondean todos los números al entero más cercano. Si el error de redondeo se puede representar como una variable aleatoria $\mathcal{U}(-\frac{1}{2}, \frac{1}{2})$ y se suman 12 números, ¿cuál es la probabilidad de que el error de redondeo exceda 1?

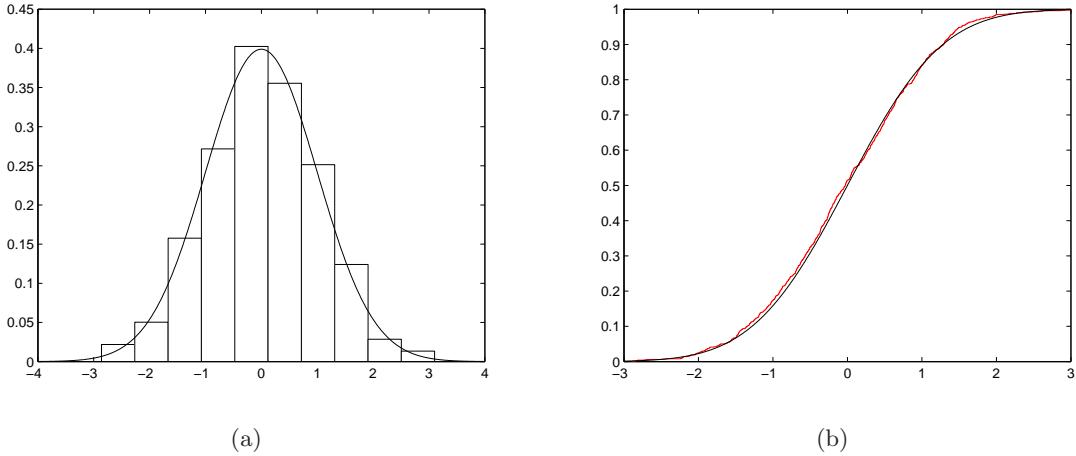


Figura 5: (a) Comparación entre un histograma de 1000 valores, cada uno de los cuales es la suma de 12 variables uniformes $\mathcal{U} [-\frac{1}{2}, \frac{1}{2}]$, y la función densidad normal; (b) Comparación entre la función de distribución empírica correspondiente a 1000 valores de la suma de 12 uniformes $\mathcal{U} [-\frac{1}{2}, \frac{1}{2}]$ y la función de distribución normal. El ajuste es sorprendentemente bueno, especialmente si se tiene en cuenta que 12 no se considera un número muy grande.

Solución: El error de redondeo cometido al sumar 12 números se representa por la suma $\sum_{i=1}^{12} X_i$ de 12 variables aleatorias independientes X_1, \dots, X_{12} cada una con distribución uniforme sobre el intervalo $(-\frac{1}{2}, \frac{1}{2})$. El error de redondeo excede 1 si y solamente si $\left| \sum_{i=1}^{12} X_i \right| > 1$. Puesto que $\mathbb{E}[X_i] = 0$ y $\mathbb{V}(X_i) = \frac{1}{12}$ de acuerdo con el teorema central del límite tenemos que la distribución de

$$\frac{\sum_{i=1}^{12} X_i - 12\mathbb{E}[X_i]}{\sqrt{12\mathbb{V}(X_i)}} = \sum_{i=1}^{12} X_i$$

se puede aproximar por la distribución normal estándar. En consecuencia,

$$\begin{aligned} \mathbb{P} \left(\left| \sum_{i=1}^{12} X_i \right| > 1 \right) &= 1 - \mathbb{P} \left(\left| \sum_{i=1}^{12} X_i \right| \leq 1 \right) \approx 1 - (\Phi(1) - \Phi(-1)) \\ &= 1 - (2\Phi(1) - 1) = 2 - 2\Phi(1) = 0.3173... \end{aligned}$$

□

Ejemplo 3.5 (Suma de exponenciales). La suma S_n de n variables aleatorias independientes exponenciales de intensidad $\lambda = 1$ obedece a una distribución gamma, $S_n \sim \Gamma(n, 1)$. En la siguiente figura se comparan, para distintos valores de n , la función de distribución de la suma estandarizada $\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}(S_n)}}$ con la función de distribución normal estándar.

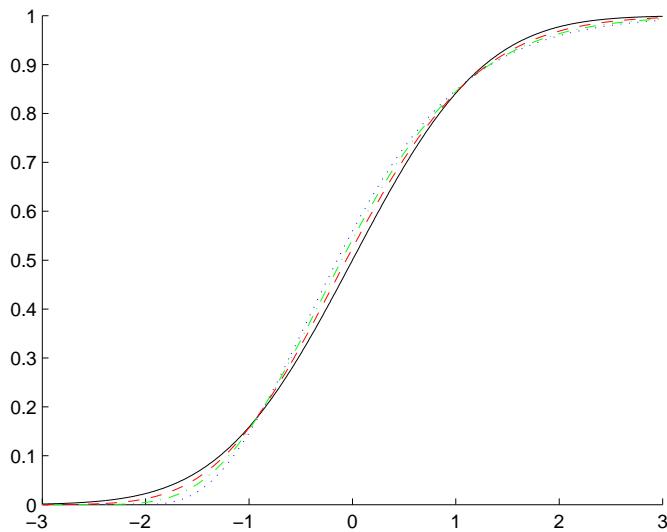


Figura 6: La normal estándar (sólida) y las funciones de distribución de las variables $\Gamma(n, 1)$ estandarizadas para $n = 5$ (punteada), $n = 10$ (quebrada y punteada) y $n = 30$ (quebrada).

Ejemplo 3.6. La distribución de Poisson de media λ se puede aproximar por la normal para valores grandes de λ : si $N \sim \text{Poisson}(\lambda)$, entonces

$$\frac{N - \lambda}{\sqrt{\lambda}} \approx \mathcal{N}(0, 1).$$

□

Ejemplo 3.7. Si la emisión de una cierta clase de partículas obedece a un proceso de Poisson de intensidad 900 por hora, ¿cuál es la probabilidad de que se emitan más de 950 partículas en una hora determinada?

Solución: Sea N una variable Poisson de media 900. Calculamos $\mathbb{P}(N > 950)$ estandarizando

$$\mathbb{P}(N > 950) = \mathbb{P}\left(\frac{N - 900}{\sqrt{900}} > \frac{950 - 900}{\sqrt{900}}\right) \approx 1 - \Phi\left(\frac{5}{3}\right) = 0.04779.$$

□

Ejemplo 3.8. El tiempo de vida de una batería es una variable aleatoria de media 40 horas y desvío 20 horas. Una batería se usa hasta que falla, momento en el cual se la reemplaza por

una nueva. Suponiendo que se dispone de un stock de 25 baterías, cuyos tiempos de vida son independientes, aproximar la probabilidad de que pueda obtenerse un uso superior a las 1100 horas.

Solución: Si ponemos X_i para denotar el tiempo de vida de la i -ésima batería puesta en uso, lo que buscamos es el valor de $p = \mathbb{P}(X_1 + \dots + X_{25} > 1000)$, que puede aproximarse de la siguiente manera:

$$p = \mathbb{P}\left(\frac{\sum_{i=1}^{25} X_i - 1000}{20\sqrt{25}} > \frac{1100 - 1000}{20\sqrt{25}}\right) \approx 1 - \Phi(1) = 0.1587.$$

□

Ejemplo 3.9. El peso W (en toneladas) que puede resistir un puente sin sufrir daños estructurales es una variable aleatoria con distribución normal de media 1400 y desvío 100. El peso (en toneladas) de cada camión de arena es una variable aleatoria de media 22 y desvío 0.25. Calcular la probabilidad de que ocurran daños estructurales cuando hay 64 camiones de arena sobre el tablero del puente.

Solución: Ocurren daños estructurales cuando la suma de los pesos de los 64 camiones, X_1, \dots, X_{64} , supera al peso W . Por el teorema central del límite, la distribución de la suma $\sum_{i=1}^{64} X_i$ es aproximadamente una normal de media 1408 y desvío 2. En consecuencia, $W - \sum_{i=1}^{64} X_i$ se distribuye (aproximadamente) como una normal de media $1400 - 1408 = -8$ y varianza $10000 + 4 = 10004$. Por lo tanto,

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{64} X_i > W\right) &= \mathbb{P}\left(W - \sum_{i=1}^{64} X_i < 0\right) = \mathbb{P}\left(\frac{W - \sum_{i=1}^{64} X_i + 8}{\sqrt{10004}} < \frac{8}{\sqrt{10004}}\right) \\ &\approx \Phi(0.07998...) = 0.5318... \end{aligned}$$

□

Ejercicios adicionales

1. Un astronauta deberá permanecer 435 días en el espacio y tiene que optar entre dos alternativas. Utilizar 36 tanques de oxígeno de tipo A o 49 tanques de oxígeno de tipo B . Cada tanque de oxígeno de tipo A tiene un rendimiento de media 12 días y desvío $1/4$. Cada tanque de oxígeno de tipo B tiene un rendimiento de media de 8,75 días y desvío $25/28$. ¿Qué alternativa es la más conveniente?
2. 432 números se redondean al entero más cercano y se suman. Suponiendo que los errores individuales de redondeo se distribuyen uniformemente sobre el intervalo $(-0.5, 0.5)$, aproximar la probabilidad de que la suma de los números redondeados difiera de la suma exacta en más de 6.
3. Dos aerolíneas A y B que ofrecen idéntico servicio para viajar de Buenos Aires a San Pablo compiten por la misma población de 400 clientes, cada uno de los cuales elige una aerolínea al azar. ¿Cuál es la probabilidad de que la línea A tenga más clientes que sus 210 asientos?

4. Distribuciones relacionadas con la Normal

En esta sección se presentan tres distribuciones de probabilidad relacionadas con la distribución normal: las distribuciones χ^2 , t y F . Esas distribuciones aparecen en muchos problemas estadísticos.

4.1. χ^2 (chi-cuadrado)

Definición 4.1 (Distribución chi-cuadrado con un grado de libertad). Si Z es una variable aleatoria con distribución normal estándar, la distribución de $U = Z^2$ se llama la distribución *chi-cuadrado con 1 grado de libertad*.

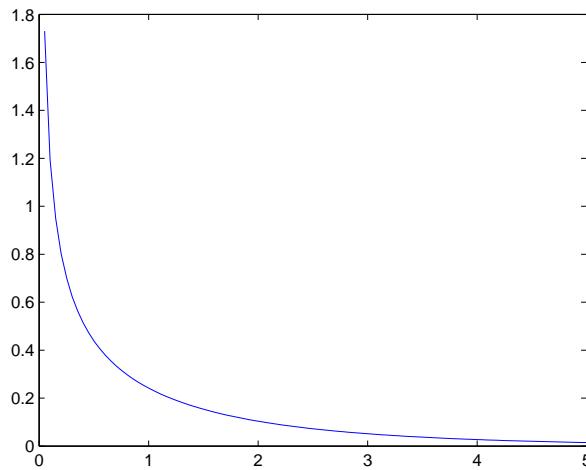


Figura 7: Gráfico de la función densidad de probabilidad de la distribución χ_1^2 .

Caracterización de la distribución χ_1^2 . La función de distribución de la variable $U = Z^2$ es $F_U(u) = \mathbb{P}(Z^2 \leq u)$, donde Z es $\mathcal{N}(0, 1)$. Para cada $u > 0$, vale que

$$F(x) = \mathbb{P}(Z^2 \leq u) = \mathbb{P}(|Z| \leq \sqrt{u}) = \mathbb{P}(-\sqrt{u} \leq Z \leq \sqrt{u}) = \int_{-\sqrt{u}}^{\sqrt{u}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Usando el teorema fundamental del cálculo integral y la regla de la cadena obtenemos que para cada $u > 0$ vale que

$$\begin{aligned} f_U(u) &= \frac{d}{du} F_U(u) = \frac{d}{du} \int_{-\sqrt{u}}^{\sqrt{u}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \left(e^{-(\sqrt{u})^2/2} \frac{d}{du} (\sqrt{u}) - e^{-(\sqrt{u})^2/2} \frac{d}{du} (-\sqrt{u}) \right) \\ &= \frac{1}{\sqrt{2\pi}} \left(e^{-u/2} \frac{1}{2\sqrt{u}} + e^{-u/2} \frac{1}{2\sqrt{u}} \right) = \frac{1}{\sqrt{2\pi}} \left(e^{-u/2} \frac{1}{\sqrt{u}} \right) \\ &= \frac{(1/2)^{1/2}}{\sqrt{\pi}} \left(u^{-1/2} e^{-(1/2)u} \right) = \frac{(1/2)^{1/2}}{\sqrt{\pi}} u^{1/2-1} e^{-(1/2)u}. \end{aligned} \tag{30}$$

La última expresión que aparece en el lado derecho de la identidad (30) es la expresión de la densidad de la distribución $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$. Por lo tanto,

$$\chi_1^2 = \Gamma\left(\frac{1}{2}, \frac{1}{2}\right).$$

Nota Bene. Notar que si $X \sim \mathcal{N}(\mu, \sigma^2)$, entonces $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$, y por lo tanto $\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi_1^2$.

Definición 4.2 (Distribución chi-cuadrado). Si U_1, U_2, \dots, U_n son variables aleatorias independientes, cada una con distribución χ_1^2 , la distribución de $V = \sum_{i=1}^n U_i$ se llama distribución *chi-cuadrado con n grados de libertad* y se denota χ_n^2 .

Caracterización de la distribución chi-cuadrado. La distribución χ_n^2 es un caso particular de la distribución Gamma. Más precisamente,

$$\chi_n^2 = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right).$$

Basta recordar que la suma de variables Γ i.i.d. también es Γ . En particular, la función densidad de V es

$$f_V(v) = \frac{(1/2)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} v^{\frac{n}{2}-1} e^{-\frac{1}{2}v} \mathbf{1}\{v > 0\}.$$

□

Nota Bene. La distribución χ_n^2 no es simétrica.

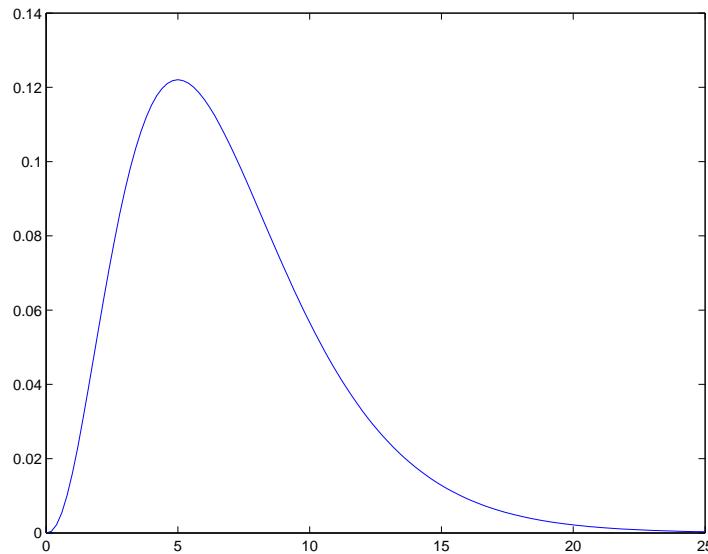


Figura 8: Gráfico de la función densidad de probabilidad de la distribución χ_7^2 .

4.2. t de Student

Definición 4.3 (La distribución t de Student). Sean Z y U variables aleatorias independientes con distribuciones $\mathcal{N}(0, 1)$ y χ_n^2 , respectivamente. La distribución de la variable

$$T = \frac{Z}{\sqrt{U/n}}$$

se llama distribución t de Student con n grados de libertad y se denota mediante t_n .

La función densidad de la t de Student con n grados de libertad es

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

La fórmula de la densidad se obtiene por los métodos estándar desarrollados en las notas sobre transformaciones de variables.

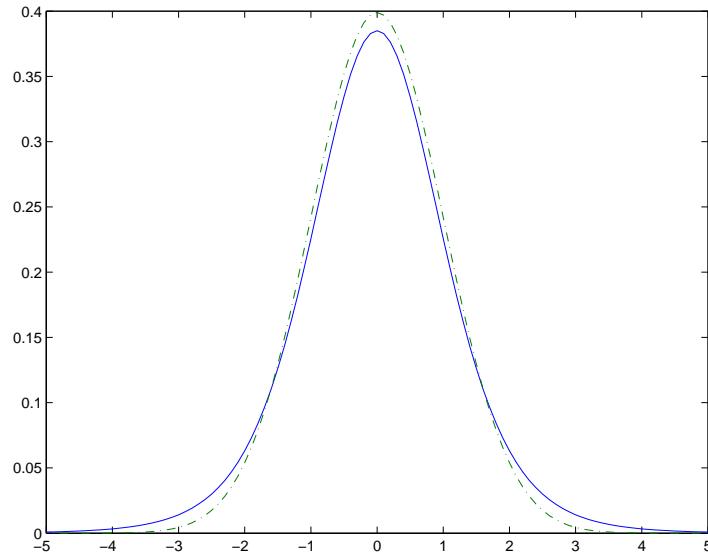


Figura 9: Comparación de la función densidad de probabilidad de una distribución t_7 (línea sólida) con la de la distribución $\mathcal{N}(0, 1)$ (línea punteada).

Observación 4.4. Notar que la densidad de t_n es simétrica respecto del origen. Cuando la cantidad de grados de libertad, n , es grande la distribución t_n se aproxima a la la distribución $\mathcal{N}(0, 1)$; de hecho para más de 20 o 30 grados de libertad, las distribuciones son muy cercanas.

4.3. F de Fisher

Definición 4.5 (Distribución F). Sean U y V variables aleatorias independientes con distribuciones χ_m^2 y χ_n^2 , respectivamente. La distribución de la variable

$$W = \frac{U/m}{V/n}$$

se llama distribución F con m y n grados de libertad y se denota por $F_{m,n}$.

La función densidad de W es

$$f_W(w) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} w^{\frac{m}{2}-1} \left(1 + \frac{m}{n}w\right)^{-\frac{m+n}{2}} \mathbf{1}\{w \geq 0\}.$$

W es el cociente de dos variables aleatorias independientes, y su densidad se obtiene usando los métodos estándar desarrollados en las notas sobre transformaciones de variables.

Nota Bene. Se puede mostrar que, para $n > 2$, $\mathbb{E}[W] = n/(n - 2)$. De las definiciones de las distribuciones t y F , se deduce que el cuadrado de una variable aleatoria t_n se distribuye como una $F_{1,n}$.

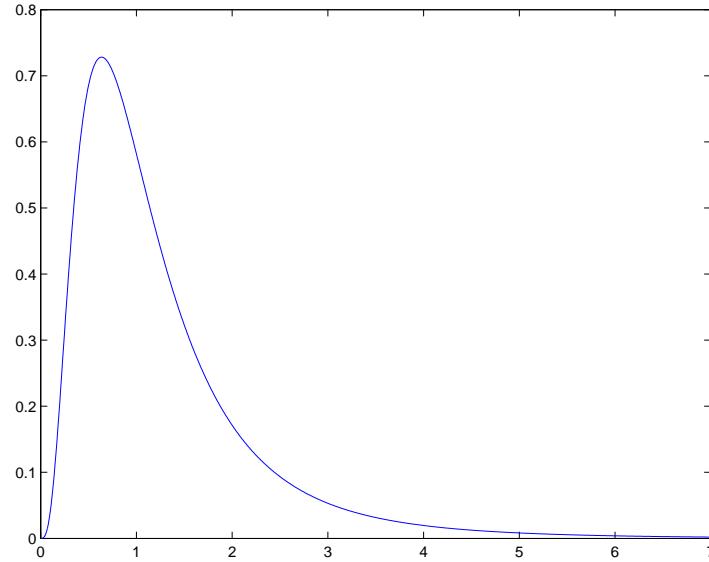


Figura 10: Gráfico típico de la función densidad de probabilidad de una distribución F .

¿Cómo usar las tablas de las distribuciones F ? Para cada $\alpha \in (0, 1)$, sea $F_{\alpha,m,n}$ el punto del semieje positivo de las abscisas a cuya derecha la distribución $F_{m,n}$ acumula una probabilidad α :

$$\mathbb{P}(F_{m,n} > F_{\alpha,m,n}) = \alpha.$$

Observación 4.6. Notar que de las igualdades

$$\alpha = \mathbb{P}\left(\frac{U/m}{V/n} > F_{\alpha,m,n}\right) = \mathbb{P}\left(\frac{V/n}{U/m} < \frac{1}{F_{\alpha,m,n}}\right) = 1 - \mathbb{P}\left(\frac{V/n}{U/m} \geq \frac{1}{F_{\alpha,m,n}}\right)$$

se deduce que

$$F_{1-\alpha,n,m} = \frac{1}{F_{\alpha,m,n}}. \quad (31)$$

En los manuales de estadística se pueden consultar las tablas de los valores $F_{\alpha,m,n}$ para diferentes valores de m, n y $\alpha \in \{0.01, 0.05\}$. Por ejemplo, según la tabla que tengo a mi disposición¹

$$\mathbb{P}(F_{9,9} > 3.18) = 0.05 \quad \text{y} \quad \mathbb{P}(F_{9,9} > 5.35) = 0.01$$

Usando esa información queremos hallar valores ϕ_1 y ϕ_2 tales que

$$\mathbb{P}(F_{9,9} > \phi_2) = 0.025 \quad \text{y} \quad \mathbb{P}(F_{9,9} < \phi_1) = 0.025.$$

El valor de ϕ_2 se obtiene por interpolación líneal entre los dos puntos dados en la tabla: $A = (3.18, 0.05)$ y $B = (5.35, 0.01)$. La ecuación de la recta que pasa por ellos es $y - 0.01 = -\frac{0.04}{2.17}(x - 5.35)$. En consecuencia, ϕ_2 será la solución de la ecuación $0.025 - 0.01 = -\frac{0.04}{2.17}(\phi_2 - 5.35)$. Esto es, $\phi_2 = 4.5362$.

El valor de ϕ_1 se obtiene observando que la ecuación $\mathbb{P}(F_{9,9} < \phi_1) = 0.025$ es equivalente a la ecuación $\mathbb{P}(1/F_{9,9} > 1/\phi_1) = 0.025$. Por definición, la distribución de $1/F_{9,9}$ coincide con la de $F_{9,9}$. En consecuencia, ϕ_1 debe satisfacer la ecuación $\mathbb{P}(F_{9,9} > 1/\phi_1) = 0.025$. Por lo tanto, $\phi_1 = 1/4.5362 = 0.2204$.

5. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Cramer, H.: Métodos matemáticos de estadística. Aguilar, Madrid. (1970)
2. Durrett R.: Probability. Theory and Examples. Duxbury Press, Belmont. (1996)
3. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons, New York. (1968)
4. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 2. John Wiley & Sons, New York. (1971)
5. Hoel P. G.: Introducción a la estadística matemática. Ariel, Barcelona. (1980)
6. Piskunov, N.: Cálculo diferencial e integral, tomo I. Mir, Moscú (1983)
7. Rice, J. A.: Mathematical Statistics and Data Analysis. Duxbury Press, Belmont. (1995)
8. Ross, S. M: Introduction to Probability and Statistics for Engineers and Scientists. Elsevier Academic Press, San Diego. (2004)
9. Ross, S.: Introduction to Probability Models. Academic Press, San Diego. (2007)

¹Introducción a la estadística matemática. Ariel, Barcelona. (1980).

Estimadores puntuales
(Borradores, Curso 23)

Sebastian Grynberg

20-22 de mayo de 2013



*La libertad de los pueblos no consiste en palabras,
ni debe existir en los papeles solamente. (...)
Si deseamos que los pueblos sean libres,
observemos religiosamente el sagrado dogma de la igualdad.
(Mariano Moreno)*

Índice

1. Introducción	2
1.1. Nociones y presupuestos básicos	2
1.2. Algunas familias paramétricas	3
2. Estimadores	4
2.1. Error cuadrático medio, sesgo y varianza	5
2.2. Comparación de estimadores	7
2.3. Consistencia	9
3. Método de máxima verosimilitud	10
3.1. Estimador de máxima verosimilitud (emv)	10
3.2. Cálculo del emv para familias regulares	12
3.2.1. Familias exponenciales	17
3.2.2. Malas noticias!	19
3.3. Cálculo del emv para familias no regulares	20
3.4. Principio de invariancia	22
4. Bibliografía consultada	23

1. Introducción

1.1. Nociones y presupuestos básicos

Definición 1.1 (Muestra aleatoria). Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad y $X : \Omega \rightarrow \mathbb{R}$ una variable aleatoria. Una *muestra aleatoria de volumen n* de la variable aleatoria X es una sucesión X_1, \dots, X_n de variables aleatorias independientes cada una con la misma distribución de X .

Modelos paramétricos. En todo lo que sigue vamos a suponer que

1. La función de distribución de la variable aleatoria X es *desconocida parcialmente*: se sabe que $F(x) = \mathbb{P}(X \leq x)$ pertenece a una familia, \mathcal{F} , de distribuciones conocidas que dependen de un parámetro θ desconocido: $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.
2. El conjunto paramétrico, Θ , es no vacío y está contenido en \mathbb{R}^d .
3. Las distribuciones de la familia \mathcal{F} son *distinguibles*: $F_{\theta_1} \neq F_{\theta_2}$ cuando $\theta_1 \neq \theta_2$.
4. Las distribuciones de la familia \mathcal{F} tienen “densidad”. Si se trata de una familia de *distribuciones continuas* esto significa que para cada $\theta \in \Theta$, existe una función densidad de probabilidades (f.d.p.) $f(x|\theta)$ tal que $\frac{d}{dx}F_\theta(x) = f(x|\theta)$. Si se trata de una familia de *distribuciones discretas* esto significa que para cada $\theta \in \Theta$, existe una función de probabilidad (f.p.) $f(x|\theta)$ tal que $F_\theta(x) - F_\theta(x-) = f(x|\theta)$.
5. Es posible conseguir muestras aleatorias de la variable X del volumen que se deseé.

Nota Bene. De los presupuestos básicos adoptados resulta que los modelos paramétricos adoptan la forma

$$\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\},$$

donde θ es un parámetro desconocido que puede tomar valores en un espacio paramétrico $\Theta \subset \mathbb{R}^d$.

1.2. Algunas familias paramétricas

Repasamos algunas de las familias de distribuciones que se utilizan comúnmente en el análisis de datos en problemas prácticos.

1. Familia Normal, $\mathcal{N}(\mu, \sigma^2)$. Decimos que X tiene distribución normal de parámetros $\mu \in \mathbb{R}$ y $\sigma^2 > 0$ cuando la f.d.p. de X está dada por

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

Vale que $\mathbb{E}[X] = \mu$ y $\mathbb{V}(X) = \sigma^2$.

2. Familia Gamma, $\Gamma(\nu, \lambda)$. Decimos que X tiene distribución gamma de parámetros $\nu > 0$ y $\lambda > 0$ cuando la f.d.p. de X está dada por

$$f(x|\nu, \lambda) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x} \mathbf{1}\{x \geq 0\},$$

donde $\Gamma(\nu) := \int_0^\infty x^{\nu-1} e^{-x} dx$. Vale que $\mathbb{E}[X] = \nu/\lambda$ y $\mathbb{V}(X) = \nu/\lambda^2$.

Casos particulares de las familias Gamma son las familias exponenciales $Exp(\lambda) = \Gamma(1, \lambda)$ y las familias chi cuadrado $\chi_\nu^2 = \Gamma(\nu/2, 1/2)$.

3. Familia Beta, $\beta(\nu_1, \nu_2)$. Decimos que X tiene distribución beta de parámetros $\nu_1 > 0$ y $\nu_2 > 0$ cuando la f.d.p. de X está dada por

$$f(x|\nu_1, \nu_2) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1} (1-x)^{\nu_2-1} \mathbf{1}\{0 < x < 1\}.$$

Vale que

$$\mathbb{E}[X] = \frac{\nu_1}{\nu_1 + \nu_2} \quad \text{y} \quad \mathbb{V}(X) = \frac{\nu_1 \nu_2}{(\nu_1 + \nu_2)^2 (\nu_1 + \nu_2 + 1)}.$$

Notar que cuando los parámetros ν_1 y ν_2 son números naturales se tiene que

$$\frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} = \frac{(\nu_1 + \nu_2 - 1)!}{(\nu_1 - 1)!(\nu_2 - 1)!} = (\nu_1 + \nu_2 - 1) \binom{\nu_1 + \nu_2 - 2}{\nu_1 - 1}.$$

La distribución $\beta(\nu_1, \nu_2)$ se puede obtener como la distribución del cociente $X_1/(X_1 + X_2)$ donde $X_1 \sim \Gamma(\nu_1, 1)$ y $X_2 \sim \Gamma(\nu_2, 1)$.

Notar que $\beta(1, 1) = \mathcal{U}(0, 1)$.

4. Familia Binomial, Binomial(n, p). Decimos que X tiene distribución Binomial de parámetros $n \in \mathbb{N}$ y $0 < p < 1$ cuando su f.p. está dada por

$$f(x|n, p) = \binom{n}{x} (1-p)^{n-x} p^x, \quad x = 0, 1, \dots, n.$$

Vale que $\mathbb{E}[X] = np$ y $\mathbb{V}(X) = np(1-p)$.

5. Familia Pascal, Pascal(n, p). Decimos que X tiene distribución Pascal de parámetros $n \in \mathbb{N}$ y $0 < p < 1$ cuando su f.p. está dada por

$$f(x|n, p) = \binom{x-1}{n-1} p^n (1-p)^{x-n}, \quad x = n, n+1, \dots$$

Vale que $\mathbb{E}[X] = n/p$ y $\mathbb{V}(X) = n(1-p)/p^2$.

6. Familia Poisson, Poisson(λ). Decimos que X tiene distribución Poisson de parámetro $\lambda > 0$ cuando su f.p. está dada por

$$f(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, \dots$$

Vale que $\mathbb{E}[X] = \lambda$ y $\mathbb{V}(X) = \lambda$.

2. Estimadores

El punto de partida de la investigación estadística está constituido por una muestra aleatoria, $\mathbf{X} = (X_1, \dots, X_n)$, de la distribución desconocida F perteneciente a una familia paramétrica de distribuciones $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ ¹. Como las distribuciones de la familia \mathcal{F} son *distinguibles* lo que se quiere saber es cuál es el parámetro $\theta \in \Theta$ que corresponde a la distribución F . En otras palabras, se quiere hallar $\theta \in \Theta$ tal que $F = F_\theta$.

Formalmente, “cualquier” función, $\hat{\theta} := \hat{\theta}(\mathbf{X})$, de la muestra aleatoria \mathbf{X} que no depende de parámetros desconocidos se denomina una *estadística*.

Ejemplo 2.1. Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de la variable aleatoria X con función de distribución F_θ . Ejemplos de estadísticas son

$$(i) \quad X_{(1)} = \min(X_1, \dots, X_n),$$

$$(ii) \quad X_{(n)} = \max(X_1, \dots, X_n),$$

$$(iii) \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

$$(iv) \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

¹**Notación.** Si \mathcal{F} es una familia de distribuciones F_θ con “densidades” $f(x|\theta)$, $\theta \in \Theta$, escribimos

$$\mathbb{P}_\theta(X \in A) = \int_A f(x|\theta) dx \quad \text{y} \quad \mathbb{E}_\theta[r(X)] = \int r(x)f(x|\theta)dx$$

El subíndice θ indica que la probabilidad o la esperanza es con respecto a $f(x|\theta)$. Similarmente, escribimos \mathbb{V}_θ para la varianza.

En (i) y (ii), $\min(\cdot)$ y $\max(\cdot)$ denotan, respectivamente, el mínimo y el máximo muestrales observados. Por otro lado, \bar{X} y $\hat{\sigma}^2$ denotan, respectivamente, la media y la varianza muestrales. \square

Cualquier estadística que asuma valores en el conjunto paramétrico Θ de la familia de distribuciones \mathcal{F} se denomina un *estimador puntual* para θ . El adjetivo puntual está puesto para distinguirla de las *estimaciones por intervalo* que veremos más adelante.

En muchas situaciones lo que interesa es estimar una función $g(\theta)$. Por ejemplo, cuando se considera una muestra aleatoria \mathbf{X} de una variable $X \sim \mathcal{N}(\mu, \sigma^2)$ donde μ y σ^2 son desconocidos entonces $\theta = (\mu, \sigma^2)$ y el conjunto de parámetros es $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R} \text{ y } \sigma^2 > 0\}$. Si el objetivo es estimar solamente μ , entonces $g(\theta) = \mu$.

Definición 2.2. Cualquier estadística que solamente asuma valores en el conjunto de los posibles valores de $g(\theta)$ es un *estimador para* $g(\theta)$.

Uno de los grandes problemas de la estadística es construir estimadores razonables para el parámetro desconocido θ o para una función $g(\theta)$. Existen diversos métodos para elegir entre todos los estimadores posibles de θ . Cada elección particular del estimador depende de ciertas propiedades que se consideran “deseables” para la estimación.

2.1. Error cuadrático medio, sesgo y varianza

Uno de los procedimientos más usados para evaluar el desempeño de un estimador es considerar su error cuadrático medio. Esta noción permite precisar el sentido que se le otorga a los enunciados del tipo “el estimador puntual $\hat{\theta} = \hat{\theta}(\mathbf{X})$ está próximo de θ ”.

Definición 2.3 (Error cuadrático medio). El *error cuadrático medio* (ECM) de un estimador $\hat{\theta}$ para el parámetro θ se define por

$$\text{ECM}(\hat{\theta}) = \mathbb{E}_\theta \left[(\hat{\theta} - \theta)^2 \right]. \quad (1)$$

El ECM se puede descomponer de la siguiente manera²

$$\mathbb{E}_\theta \left[(\hat{\theta} - \theta)^2 \right] = \mathbb{V}_\theta(\hat{\theta}) + \mathbb{B}_\theta^2(\hat{\theta}), \quad (2)$$

donde $\mathbb{B}_\theta(\hat{\theta}) := \mathbb{E}_\theta[\hat{\theta}] - \theta$ es el llamado *sesgo* del estimador. El primer término de la descomposición (2) describe la “variabilidad” del estimador, y el segundo el “error sistemático”: $\mathbb{E}_\theta[\hat{\theta}]$ describe alrededor de qué valor fluctúa $\hat{\theta}$ y $\mathbb{V}_\theta(\hat{\theta})$ mide cuánto fluctúa.

²La descomposición (2) se obtiene escribiendo $\hat{\theta} - \theta$ en la forma $(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]) + (\mathbb{E}_\theta[\hat{\theta}] - \theta)$. Desarrollando cuadrados obtenemos $(\hat{\theta} - \theta)^2 = (\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])(\mathbb{E}_\theta[\hat{\theta}] - \theta) + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2$. El resultado se obtiene observando que la esperanza \mathbb{E}_θ de los términos cruzados $(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])(\mathbb{E}_\theta[\hat{\theta}] - \theta)$ es igual a 0:

$$\begin{aligned} \mathbb{E}_\theta \left[(\hat{\theta} - \theta)^2 \right] &= \mathbb{E}_\theta \left[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])(\mathbb{E}_\theta[\hat{\theta}] - \theta) + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2 \right] \\ &= \mathbb{E}_\theta \left[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2 \right] + 0 + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2 = \mathbb{V}_\theta(\hat{\theta}) + \mathbb{B}_\theta^2(\hat{\theta}). \end{aligned}$$

Definición 2.4 (Estimadores insesgados). Diremos que un estimador $\hat{\theta}$ es *insesgado* para el parámetro θ si

$$\mathbb{E}_\theta[\hat{\theta}] = \theta.$$

para todo $\theta \in \Theta$, o sea $\mathbb{B}_\theta(\hat{\theta}) \equiv 0$. Si $\lim_{n \rightarrow \infty} \mathbb{B}_\theta[\hat{\theta}] = 0$ para todo $\theta \in \Theta$, diremos que el estimador $\hat{\theta}$ es *asintóticamente insesgado* para θ .

Nota Bene. En el caso en que $\hat{\theta}$ es un estimador insesgado para θ , tenemos que

$$\text{ECM}(\hat{\theta}) = \mathbb{V}_\theta(\hat{\theta}),$$

o sea, el error cuadrático medio de $\hat{\theta}$ se reduce a su varianza.

Nota Bene. Una consecuencia destacable de la descomposición (2) para grandes muestras ($n \gg 1$) es la siguiente: si a medida que se aumenta el volumen de la muestra, el sesgo y la varianza del estimador $\hat{\theta}$ tienden a cero, entonces, el estimador $\hat{\theta}$ converge en media cuadrática al verdadero valor del parámetro θ .

Ejemplo 2.5 (Estimación de media). Sea $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ una familia de distribuciones. Para cada $\theta \in \Theta$ designemos mediante $\mu(\theta)$ y $\sigma^2(\theta)$ la media y la varianza correspondientes a la distribución F_θ , respectivamente. Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de alguna distribución perteneciente a \mathcal{F} . Denotemos mediante \bar{X} el promedio de la muestra:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

En lo que sigue vamos a suponer que para cada $\theta \in \Theta$, $\mu(\theta) \in \mathbb{R}$ y $\sigma^2(\theta) < \infty$. Si la muestra aleatoria proviene de la distribución F_θ , tenemos que

$$\mathbb{E}_\theta[\bar{X}] = \mathbb{E}_\theta \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta[X_i] = \mu(\theta).$$

Por lo tanto \bar{X} es un estimador insesgado para $\mu(\theta)$ y su error cuadrático medio al estimar $\mu(\theta)$ es

$$\text{ECM}(\bar{X}) = \mathbb{V}_\theta(\bar{X}) = \mathbb{V}_\theta \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}_\theta[X_i] = \frac{1}{n} \sigma^2(\theta).$$

□

Ejemplo 2.6 (Estimación de varianza). Sea $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ una familia de distribuciones. Para cada $\theta \in \Theta$ designemos mediante $\mu(\theta)$ y $\sigma^2(\theta)$ la media y la varianza correspondientes a la distribución F_θ , respectivamente, a las que supondremos finitas. Sea X_1, \dots, X_n una muestra aleatoria de alguna distribución perteneciente a \mathcal{F} . Sean \bar{X} y $\hat{\sigma}^2$ la media y la varianza muestrales definidas en el Ejemplo 2.1:

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i \quad \text{y} \quad \hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Para analizar el sesgo de la varianza muestral conviene descomponerla de la siguiente manera:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu(\theta))^2 - (\bar{X} - \mu(\theta))^2, \quad (3)$$

cualquiera sea $\theta \in \Theta$.³ Si la muestra aleatoria, X_1, \dots, X_n , proviene de la distribución F_θ , al tomar esperanzas en ambos lados de (3) se obtiene

$$\begin{aligned} \mathbb{E}_\theta[\hat{\sigma}^2] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta[(X_i - \mu(\theta))^2] - \mathbb{E}_\theta[(\bar{X} - \mu(\theta))^2] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{V}_\theta(X_i) - \mathbb{V}_\theta(\bar{X}). \end{aligned} \quad (4)$$

Según el Ejemplo 2.5 \bar{X} es un estimador insesgado para la media $\mu(\theta)$ y su varianza vale $\mathbb{V}_\theta(\bar{X}) = \frac{1}{n}\sigma^2(\theta)$, en consecuencia,

$$\mathbb{E}_\theta[\hat{\sigma}^2] = \frac{1}{n} \sum_{i=1}^n \mathbb{V}_\theta(X_i) - \mathbb{V}_\theta(\bar{X}) = \sigma^2(\theta) - \frac{1}{n}\sigma^2(\theta) = \frac{n-1}{n}\sigma^2(\theta). \quad (5)$$

Esto demuestra que $\hat{\sigma}^2$ no es un estimador insesgado para la varianza $\sigma^2(\theta)$. La identidad $\mathbb{E}_\theta[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2(\theta)$ significa que si tomamos repetidas muestras de tamaño n y se promedian las varianzas muestrales resultantes, el promedio no se aproximará a la verdadera varianza, sino que de modo sistemático el valor será más pequeño debido al factor $(n-1)/n$. Este factor adquiere importancia en las muestras pequeñas. Si $n \rightarrow \infty$, el factor $(n-1)/n \rightarrow 1$ lo que demuestra que $\hat{\sigma}^2$ es un estimador asintóticamente insesgado para la varianza $\sigma^2(\theta)$.

Para eliminar el sesgo en $\hat{\sigma}^2$, basta multiplicar $\hat{\sigma}^2$ por $\frac{n}{n-1}$. De (5) sigue que

$$S^2 := \frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (6)$$

es un estimador insesgado para la varianza. □

2.2. Comparación de estimadores

El error cuadrático medio puede usarse para comparar estimadores. Diremos que $\hat{\theta}_1$ es mejor que $\hat{\theta}_2$ si

$$\text{ECM}(\hat{\theta}_1) \leq \text{ECM}(\hat{\theta}_2), \quad (7)$$

para todo θ , con desigualdad estricta para al menos un valor de θ . En tal caso, el estimador $\hat{\theta}_2$ se dice *inadmissible*. Si existe un estimador $\hat{\theta}^*$ tal que para todo estimador $\hat{\theta}$ de θ con $\hat{\theta} \neq \hat{\theta}^*$

$$\text{ECM}(\hat{\theta}^*) \leq \text{ECM}(\hat{\theta}), \quad (8)$$

³La descomposición (3) se obtiene haciendo lo siguiente. Para cada i escribimos $(X_i - \bar{X})$ en la forma $(X_i - \mu(\theta)) - (\bar{X} - \mu(\theta))$. Desarrollando cuadrados obtenemos $(X_i - \bar{X})^2 = (X_i - \mu(\theta))^2 + (\bar{X} - \mu(\theta))^2 - 2(X_i - \mu(\theta))(\bar{X} - \mu(\theta))$. El resultado se obtiene observando que el promedio de los términos cruzados $(X_i - \mu(\theta))(\bar{X} - \mu(\theta))$ es igual a $(\bar{X} - \mu(\theta))^2$. (*Hacer la cuenta y verificarlo!*)

para todo θ , con desigualdad estricta para al menos un valor de θ , entonces $\hat{\theta}^*$ se dice *óptimo*.

Cuando la comparación se restringe a los estimadores son insesgados, el estimador óptimo, $\hat{\theta}^*$, se dice el estimador insesgado de varianza uniformemente mínima. Esta denominación resulta de observar que estimadores insesgados la relación (8) adopta la forma

$$\mathbb{V}_\theta(\hat{\theta}^*) \leq \mathbb{V}_\theta(\hat{\theta}),$$

para todo θ , con desigualdad estricta para al menos un valor de θ .

Ejemplo 2.7. Sean X_1, X_2, X_3 una muestra aleatoria de una variable aleatoria X tal que $\mathbb{E}_\theta[X] = \theta$ y $\mathbb{V}_\theta(X) = 1$. Consideremos los estimadores

$$\bar{X} = \frac{X_1 + X_2 + X_3}{3} \quad \text{y} \quad \hat{\theta} = \frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3.$$

Según el Ejemplo 2.5 $\mathbb{E}_\theta[\bar{X}] = \theta$ y $\mathbb{V}_\theta(\bar{X}) = \frac{1}{3}$. Tenemos también que

$$\mathbb{E}_\theta[\hat{\theta}] = \frac{1}{2}\mathbb{E}_\theta[X_1] + \frac{1}{4}\mathbb{E}_\theta[X_2] + \frac{1}{4}\mathbb{E}_\theta[X_3] = \frac{1}{2}\theta + \frac{1}{4}\theta + \frac{1}{4}\theta = \theta$$

y

$$\mathbb{V}_\theta(\hat{\theta}) = \frac{1}{4}\mathbb{V}_\theta(X_1) + \frac{1}{16}\mathbb{V}_\theta(X_2) + \frac{1}{16}\mathbb{V}_\theta(X_3) = \frac{1}{4} + \frac{1}{16} + \frac{1}{16} = \frac{6}{16}.$$

Como \bar{X} y $\hat{\theta}$ son insesgados, resulta que \bar{X} es mejor que $\hat{\theta}$, pues $\mathbb{V}_\theta(\bar{X}) < \mathbb{V}_\theta(\hat{\theta})$ para todo θ .

□

Ejemplo 2.8. Sea X_1, \dots, X_n una muestra aleatoria de una variable aleatoria $X \sim \mathcal{U}(0, \theta)$. Vamos a considerar $\hat{\theta}_1 = 2\bar{X}$ y $\hat{\theta}_2 = X_{(n)}$ como estimadores para θ y estudiaremos su comportamiento. Como $\mathbb{E}_\theta[X] = \theta/2$ y $\mathbb{V}_\theta(X) = \theta^2/12$, tenemos que

$$\mathbb{E}_\theta[\hat{\theta}_1] = \mathbb{E}_\theta[2\bar{X}] = \theta \quad \text{y} \quad \mathbb{V}_\theta(\hat{\theta}_1) = \frac{\theta^2}{3n}. \quad (9)$$

Por lo tanto, $\hat{\theta}_1$ es un estimador insesgado para θ . En consecuencia,

$$\text{ECM}(\hat{\theta}_1) = \mathbb{V}_\theta(\hat{\theta}_1) = \frac{\theta^2}{3n}. \quad (10)$$

Por otro lado, la función densidad de $X_{(n)}$ está dada por $f_\theta(x) = \frac{nx^{n-1}}{\theta^n} \mathbf{1}\{0 < x < \theta\}$, de donde se deduce que

$$\mathbb{E}_\theta[X_{(n)}] = \frac{n}{n+1}\theta \quad \text{y} \quad \mathbb{V}_\theta(X_{(n)}) = \frac{n\theta^2}{(n+1)^2(n+2)}. \quad (11)$$

Por lo tanto, $\hat{\theta}_2$ es un estimador asintóticamente insesgado para θ . Combinando las identidades (11) en (2), obtenemos

$$\begin{aligned} \text{ECM}(\hat{\theta}_2) &= \mathbb{V}_\theta(\hat{\theta}_2) + \mathbb{B}_\theta^2(\hat{\theta}_2) = \frac{n\theta^2}{(n+1)^2(n+2)} + \left(\frac{n}{n+1}\theta - \theta \right)^2 \\ &= \frac{n\theta^2}{(n+1)^2(n+2)} + \frac{\theta^2}{(n+1)^2} = \frac{2\theta^2}{(n+1)(n+2)}. \end{aligned} \quad (12)$$

Es fácil, pero tedioso, ver que $\text{ECM}(\hat{\theta}_2) < \text{ECM}(\hat{\theta}_1)$ para todo θ y todo n . Por lo tanto, $X_{(n)}$ es mejor que $2\bar{X}$ para todo θ y todo n .

□

2.3. Consistencia

Lo mínimo que se le puede exigir a un estimador puntual, $\hat{\theta}(X_1, \dots, X_n)$, es que, en algún sentido, se aproxime al verdadero valor del parámetro cuando el volumen de la muestra aumenta. En otras palabras, si $\theta \in \Theta$ es tal que $F = F_\theta$ y X_1, X_2, \dots es una sucesión de variables aleatorias independientes cada una con distribución F , en algún sentido, debe ocurrir que

$$\hat{\theta}(X_1, \dots, X_n) \rightarrow \theta,$$

cuando $n \rightarrow \infty$.

Por ejemplo, es deseable que el estimador $\hat{\theta}$ tenga la siguiente propiedad, llamada *consistencia débil*: para cada $\epsilon > 0$ debe cumplir que

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(|\hat{\theta}(X_1, \dots, X_n) - \theta| > \epsilon) = 0. \quad (13)$$

Más exigente, es pedirle que tenga la siguiente propiedad, llamada *consistencia fuerte*:

$$\mathbb{P}_\theta \left(\lim_{n \rightarrow \infty} \hat{\theta}(X_1, \dots, X_n) = \theta \right) = 1. \quad (14)$$

Normalidad asintótica. También se le puede pedir una propiedad similar a la del teorema central límite, llamada *normalidad asintótica*: existe $\sigma = \sigma(\theta) > 0$ tal que

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta \left(\frac{\sqrt{n}(\hat{\theta}(X_1, \dots, X_n) - \theta)}{\sigma} \leq x \right) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \quad (15)$$

Nota Bene. Los problemas de consistencia y normalidad asintótica están relacionados con las leyes de los grandes números y el teorema central de límite. El siguiente ejemplo muestra dicha relación para el caso en que se quiere estimar la media de una distribución.

Ejemplo 2.9 (Estimación de media). Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria cuya distribución pertenece a una familia $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$. Sean $\mu(\theta)$ y $\sigma^2(\theta)$ la media y la varianza correspondientes a la distribución F_θ , respectivamente. Aplicando la desigualdad de Chebychev a \bar{X} se obtiene que para cada $\epsilon > 0$

$$\mathbb{P}_\theta (|\bar{X} - \mu(\theta)| > \epsilon) \leq \frac{\mathbb{V}_\theta(\bar{X})}{\epsilon^2} = \frac{1}{n} \left(\frac{\sigma^2(\theta)}{\epsilon^2} \right) \rightarrow 0,$$

cuando $n \rightarrow \infty$.

Hasta aquí, lo único que hicimos es volver a demostrar la ley débil de los grandes números. Lo que queremos subrayar es que *en el contexto de la estimación de parámetros, la ley débil de los grandes números significa que el promedio de la muestra, \bar{X} , es un estimador débilmente consistente para la media de la distribución, $\mu(\theta)$* .

La consistencia fuerte del promedio, como estimador para la media es equivalente a la *Ley fuerte de los grandes números* que afirma que: *Si X_1, X_2, \dots es una sucesión de variables aleatorias independientes e idénticamente distribuidas y si existe $\mathbb{E}[X_i] = \mu$, entonces*

$$\mathbb{P} \left(\lim_{n \rightarrow \infty} \bar{X} = \mu \right) = 1.$$

La normalidad asintótica es equivalente al teorema central del límite. □

Nota Bene. De todas las propiedades de convergencia la consistencia débil es la mas simple, en el sentido de que puede establecerse con unas pocas herramientas técnicas. Para verificar la consistencia débil del promedio para estimar la media solamente usamos la desigualdad de Chebychev y las propiedades de la media y la varianza. El razonamiento utilizado en el Ejemplo 2.9 se puede extender un poco más allá.

Teorema 2.10. *Sea $\hat{\theta}$ un estimador de θ basado en una muestra aleatoria de volumen n . Si $\hat{\theta}$ es asintóticamente insesgado y su varianza tiende a cero, entonces $\hat{\theta}$ es débilmente consistente.*

Demostración. El resultado se obtiene usando la desigualdad de Chebychev y la identidad (2):

$$\mathbb{P}_\theta \left(|\hat{\theta} - \theta| > \epsilon \right) \leq \frac{1}{\epsilon^2} \mathbb{E}_\theta \left[(\hat{\theta} - \theta)^2 \right] = \frac{1}{\epsilon^2} \left(\mathbb{V}_\theta(\hat{\theta}) + \mathbb{B}_\theta^2(\hat{\theta}) \right) \rightarrow 0.$$

□

3. Método de máxima verosimilitud

El método de máxima verosimilitud es un “método universal” para construir estimadores puntuales. Su base intuitiva es la siguiente: *si al realizar un experimento aleatorio se observa un resultado, este debe tener alta probabilidad de ocurrir.*

Para hacer más precisa esa base intuitiva consideremos una muestra aleatoria, $\mathbf{X} = (X_1, \dots, X_n)$, de una variable aleatoria discreta X con función de probabilidad $f(x|\theta)$, $\theta \in \Theta$, donde Θ es el espacio paramétrico. La probabilidad de observar los resultados $X_1 = x_1, \dots, X_n = x_n$ se calcula del siguiente modo:

$$\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i) = \prod_{i=1}^n f(x_i|\theta). \quad (16)$$

Si los resultados observables deben tener una alta probabilidad de ocurrir y observamos que $X_1 = x_1, \dots, X_n = x_n$, entonces lo razonable sería elegir entre todos los parámetros posibles, $\theta \in \Theta$, aquél (o aquellos) que maximicen (16). En consecuencia, se podría estimar θ como el valor (o los valores) de θ que hace máxima la probabilidad $\prod_{i=1}^n f(x_i|\theta)$.

3.1. Estimador de máxima verosimilitud (emv)

Definición 3.1 (EMV). Sea X una variable aleatoria cuya distribución pertenece a la familia paramétrica $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$. Un *estimador de máxima verosimilitud* de θ , basado en los valores $\mathbf{x} = (x_1, \dots, x_n)$ de una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$, es un valor $\hat{\theta}_{mv} \in \Theta$ que maximiza la función de verosimilitud

$$L(\theta|\mathbf{x}) := \prod_{i=1}^n f(x_i|\theta), \quad (17)$$

donde, dependiendo de la naturaleza de las distribuciones de la familia \mathcal{F} , $f(x|\theta)$ es la función de probabilidad o la función densidad de probabilidades de X .

Sobre la notación. Para destacar que el valor del estimador de máxima verosimilitud depende de los valores observados, $\mathbf{x} = (x_1, \dots, x_n)$, en lugar de $\hat{\theta}_{mv}$ escribiremos $\hat{\theta}_{mv}(\mathbf{x})$:

$$\hat{\theta}_{mv} = \hat{\theta}_{mv}(\mathbf{x}) := \arg \max_{\theta \in \Theta} L(\theta|\mathbf{x}). \quad (18)$$

Ejemplo 3.2. Supongamos que tenemos una moneda que puede ser equilibrada o totalmente cargada para que salga cara. Lanzamos la moneda n veces y registramos la sucesión de caras y cecas. Con esa información queremos estimar qué clase de moneda tenemos.

Cada lanzamiento de la moneda se modela con una variable aleatoria X con distribución Bernoulli(θ), donde θ es la probabilidad de que la moneda salga cara. El espacio paramétrico es el conjunto $\Theta = \{1/2, 1\}$.

El estimador de máxima verosimilitud para θ , basado en los valores $\mathbf{x} = (x_1, \dots, x_n)$ de una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de la variable X , es el valor de $\hat{\theta}_{mv}(\mathbf{x}) \in \Theta = \{1/2, 1\}$ que maximiza la función de verosimilitud $L(\theta|\mathbf{x})$. Para encontrarlo comparamos los valores de la función de verosimilitud $L(1/2|\mathbf{x})$ y $L(1|\mathbf{x})$:

$$L(1/2|\mathbf{x}) = \prod_{i=1}^n f(x_i|1/2) = (1/2)^n, \quad L(1|\mathbf{x}) = \mathbf{1} \left\{ \sum_{i=1}^n x_i = n \right\}.$$

En consecuencia, el estimador de máxima verosimilitud para θ , basado en los valores $\mathbf{x} = (x_1, \dots, x_n)$ de una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ es

$$\hat{\theta}_{mv}(\mathbf{x}) = \frac{1}{2} \mathbf{1} \left\{ \sum_{i=1}^n x_i < n \right\} + \mathbf{1} \left\{ \sum_{i=1}^n x_i = n \right\}.$$

Por lo tanto, el estimador de máxima verosimilitud para θ basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ es

$$\hat{\theta}_{mv}(\mathbf{X}) = \frac{1}{2} \mathbf{1} \left\{ \sum_{i=1}^n X_i < n \right\} + \mathbf{1} \left\{ \sum_{i=1}^n X_i = n \right\}.$$

Por ejemplo, si en 10 lanzamientos de la moneda se observaron 10 caras, el estimador de máxima verosimilitud para θ es $\hat{\theta}_{mv} = 1$; en cambio si se observaron 8 caras y 2 cecas, el estimador de máxima verosimilitud es $\hat{\theta}_{mv} = 1/2$. \square

Ejemplo 3.3. Sea X una variable aleatoria con función densidad dada por

$$f(x|\theta) = \frac{1}{2}(1 + \theta x) \mathbf{1}\{x \in [-1, 1]\}, \quad \theta \in [-1, 1].$$

Supongamos que queremos hallar el estimador de máxima verosimilitud para θ basado en la realización de una muestra aleatoria tamaño 1, X_1 . Si se observa el valor x_1 , la función de verosimilitud adopta la forma

$$L(\theta|x_1) = \frac{1}{2}(1 + \theta x_1)$$

El gráfico de $L(\theta|x_1)$ es un segmento de recta de pendiente x_1 . Como se trata de una recta el máximo se alcanza en alguno de los extremos del intervalo $\Theta = [-1, 1]$:

1. si $x_1 < 0$, el máximo se alcanza en $\theta = -1$,

2. si $x_1 = 0$, el máximo se alcanza en cualquiera de los valores del intervalo Θ ,
3. si $x_1 > 0$, el máximo se alcanza en $\theta = 1$.

Abusando de la notación tenemos que

$$\hat{\theta}_{mv}(x_1) = -\mathbf{1}\{x_1 < 0\} + \Theta \mathbf{1}\{x_1 = 0\} + \mathbf{1}\{x_1 > 0\}.$$

Por lo tanto,

$$\hat{\theta}_{mv}(X_1) = -\mathbf{1}\{X_1 < 0\} + \Theta \mathbf{1}\{X_1 = 0\} + \mathbf{1}\{X_1 > 0\}.$$

□

Ejemplo 3.4. Sea X una variable aleatoria con función densidad dada por

$$f(x|\theta) = \frac{1}{2}(1 + \theta x)\mathbf{1}\{x \in [-1, 1]\}, \quad \theta \in [-1, 1].$$

Supongamos que una muestra aleatoria de tamaño 2 arrojó los valores $1/2$ y $1/4$ y con esa información queremos hallar el estimador de máxima verosimilitud para θ . La función de verosimilitud adopta la forma

$$L(\theta|1/2, 1/4) = \frac{1}{4} \left(1 + \theta \frac{1}{2}\right) \left(1 + \theta \frac{1}{4}\right),$$

y su gráfico es un segmento de parábola “cóncava” cuyas raíces son -4 y -2 . Por lo tanto, $\hat{\theta}_{mv}(1/2, 1/4) = 1$.

Supongamos ahora que una muestra aleatoria de tamaño 2 arrojó los valores $1/2$ y $-1/3$ y con esa información queremos hallar el estimador de máxima verosimilitud para θ . La función de verosimilitud adopta la forma

$$L(\theta|1/2, -1/3) = \frac{1}{4} \left(1 + \theta \frac{1}{2}\right) \left(1 - \theta \frac{1}{3}\right),$$

y su gráfico es un segmento de parábola “convexa” cuyas raíces son -2 y 3 . Por lo tanto, $\hat{\theta}_{mv}(1/2, -1/3) = 0.5$. □

3.2. Cálculo del emv para familias regulares

Sea $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ una familia paramétrica de distribuciones y sea $\{f(x|\theta) : \theta \in \Theta\}$ la familia de funciones de densidad (o de probabilidad) asociada. Diremos que la familia \mathcal{F} es *regular* si satisface las siguientes condiciones:

1. El conjunto paramétrico $\Theta \subset \mathbb{R}^d$ es abierto.
2. El soporte de las funciones $f(x|\theta)$ no depende del parámetro. Esto es, existe un conjunto \mathbb{S} tal que $\text{sop } f(\cdot|\theta) := \{x \in \mathbb{R} : f(x|\theta) > 0\} = \mathbb{S}$ para todo $\theta \in \Theta$.
3. Para cada $x \in \mathbb{S}$, la función $f(x|\theta)$ tiene derivadas parciales respecto de todas las componentes θ_j , $j = 1, \dots, d$.

Supongamos ahora que $\mathbf{X} = (X_1, \dots, X_n)$ es una muestra aleatoria de tamaño n de una variable aleatoria X con función de densidad (o de probabilidad) $f(x|\theta)$, $\theta \in \Theta$, perteneciente a una familia regular de distribuciones. Debido a que la familia es regular cada uno de los valores observados pertenece al soporte común de las funciones $f(x|\theta)$: $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{S}^n$. Por lo tanto, cualesquiera sean los valores observados, $\mathbf{x} = (x_1, \dots, x_n)$, vale que

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta) > 0.$$

Esto habilita a tomar logaritmos y utilizar la propiedad “el logaritmo del producto es igual a la suma de los logaritmos”. En consecuencia, para cada $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{S}^n$, la función $\log L(\theta|\mathbf{x})$ está bien definida y vale que

$$\log L(\theta|\mathbf{x}) = \log \prod_{i=1}^n f(x_i|\theta) = \sum_{i=1}^n \log f(x_i|\theta). \quad (19)$$

Como el logaritmo natural $\log(\cdot)$ es una función monótona creciente, maximizar la función de verosimilitud $L(\theta|\mathbf{x})$ será equivalente a maximizar $\log L(\theta|\mathbf{x})$. La ventaja de maximizar el logaritmo de la función de verosimilitud es que, bajo las condiciones de regularidad enunciadas previamente, los productos se convierten en sumas, aligerando considerablemente el trabajo de cómputo del EMV ya que el EMV debe verificar el sistema de ecuaciones

$$\frac{\partial \log L(\theta|\mathbf{x})}{\partial \theta_j} = 0 \quad j = 1, \dots, d. \quad (20)$$

En vista de (19) el sistema de ecuaciones (20) se transforma en

$$\sum_{i=1}^n \frac{\partial \log f(x_i|\theta)}{\partial \theta_j} = 0, \quad j = 1, \dots, d. \quad (21)$$

Por este camino llegamos al siguiente resultado que provee la herramienta adecuada para el cálculo del EMV.

Lema 3.5. Sea X una variable aleatoria con función de densidad (o de probabilidad) $f(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, perteneciente a una familia regular de distribuciones. El estimador de máxima verosimilitud de θ , basado en los valores $\mathbf{x} = (x_1, \dots, x_n)$ de una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$, es solución del siguiente sistema de ecuaciones:

$$\sum_{i=1}^n \psi_j(\theta|x_i) = 0 \quad j = 1, \dots, d, \quad (22)$$

donde, para cada $x \in \mathbb{S}$, la funciones de θ , $\psi_j(\theta|x)$, $j = 1, \dots, d$, se definen por

$$\psi_j(\theta|x) := \frac{\partial \log f(x|\theta)}{\partial \theta_j}. \quad (23)$$

Nota Bene. Por supuesto que las condiciones (22) son necesarias pero no suficientes para que θ sea un máximo. Para asegurarse que θ es un máximo deberán verificarse las condiciones de segundo orden. Además debe verificarse que no se trata de un máximo relativo sino absoluto.

Nota Bene. Si la función de densidad (o de probabilidad) $f(x|\theta)$ de la variable aleatoria X pertenece a una familia regular *uniparamétrica* de distribuciones, i.e., cuando el espacio paramétrico Θ es un subconjunto de la recta real \mathbb{R} , el sistema de ecuaciones (22) se reduce a una sola ecuación, denominada la *ecuación de verosimilitud*,

$$\sum_{i=1}^n \psi(\theta|x_i) = 0, \quad (24)$$

donde, para cada $x \in \mathbb{S}$, la función de θ , $\psi(\theta|x)$, se define por

$$\psi(\theta|x) := \frac{\partial \log f(x|\theta)}{\partial \theta}. \quad (25)$$

Ejemplo 3.6 (Distribuciones de Bernoulli). Es fácil ver que la familia de distribuciones Bernoulli(θ), $\theta \in (0, 1)$, es una familia uniparamétrica regular con funciones de probabilidad de la forma $f(x|\theta) = (1-\theta)^{1-x}\theta^x$, $x = 0, 1$. En consecuencia, para encontrar el estimador de máxima verosimilitud para θ basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ podemos usar el resultado del Lema 3.5.

En primer lugar hallamos la expresión de la función $\psi(\theta|x) = \frac{\partial \log f(x|\theta)}{\partial \theta}$. Observando que

$$\log f(x|\theta) = \log(1-\theta)^{1-x}\theta^x = (1-x)\log(1-\theta) + x\log\theta,$$

y derivando respecto de θ obtenemos

$$\psi(\theta|x) = \frac{1}{1-\theta}(x-1) + \frac{1}{\theta}x$$

Por lo tanto, la ecuación de verosimilitud (24) adopta la forma

$$\frac{1}{1-\theta} \sum_{i=1}^n (x_i - 1) + \frac{1}{\theta} \sum_{i=1}^n x_i = 0. \quad (26)$$

Un poco de álgebra muestra que para cada pareja $a \neq b$ vale que:

$$\frac{1}{1-\theta}a + \frac{1}{\theta}b = 0 \Leftrightarrow \theta = \frac{b}{b-a}. \quad (27)$$

Sigue de (27), poniendo $a = \sum_{i=1}^n (x_i - 1) = \sum_{i=1}^n x_i - n$ y $b = \sum_{i=1}^n x_i$, que la solución de la ecuación (26) es

$$\theta = \frac{1}{n} \sum_{i=1}^n x_i.$$

Con un poco más de trabajo, se puede verificar que dicha solución maximiza el logaritmo de la verosimilitud.

En resumen, si $\mathbf{x} = (x_1, \dots, x_n)$ son los valores observados de una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$, el estimador de máxima verosimilitud para θ es el promedio (o media) muestral

$$\hat{\theta}_{mv} = \hat{\theta}_{mv}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n x_i$$

Por lo tanto, el estimador de máxima verosimilitud para θ , basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de una variable con distribución Bernoulli(θ), es el promedio muestral

$$\hat{\theta}_{mv}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i. \quad (28)$$

□

Nota Bene. El estimador de máxima verosimilitud para θ , basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$, de una variable aleatoria con distribución Bernoulli(θ),

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

es una variable aleatoria. Subrayamos este hecho para que no se pierda de vista que los estimadores puntuales son funciones de la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ y por lo tanto son variables aleatorias. En el Ejemplo 3.6, el parámetro θ es la media de la distribución que produce la muestra y el estimador de máxima verosimilitud para θ es el promedio muestral. Por lo tanto, $\hat{\theta}_{mv}$ es un estimador *insesgado, consistente y asintóticamente normal*.

Nota Bene. Si la muestra aleatoria arrojó los valores 1, 1, ..., 1, es fácil ver que $\hat{\theta}_{mv} = 1$, en cambio si arrojó 0, 0, ..., 0 resulta que $\hat{\theta}_{mv} = 0$. Estos resultados también coinciden con el promedio de los valores observados. Por lo tanto, el resultado obtenido en (28) se puede extender al caso en que $\Theta = [0, 1]$.

Ejemplo 3.7 (Distribuciones de Bernoulli). Bajo el supuesto de que los valores de la secuencia

$$0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0. \quad (29)$$

fueron arrojados por una muestra aleatoria de tamaño 20 de una variable aleatoria $X \sim \text{Bernoulli}(\theta)$, el estimador de máxima verosimilitud arrojará como resultado la siguiente estimación para el parámetro θ :

$$\hat{\theta}_{mv}(0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0) = \frac{11}{20} = 0.55$$

Con esta estimación podríamos decir que la ley que produce esos valores es la distribución de Bernoulli (0.55). Por lo tanto, si queremos “reproducir” el generador de números aleatorios que produjo esos resultados, debemos simular números aleatorios con distribución de Bernoulli de parámetro 0.55. □

Ejemplo 3.8 (Distribuciones normales con varianza conocida). Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria $X \sim \mathcal{N}(\theta, \sigma^2)$, con varianza $\sigma^2 > 0$ conocida y media $\theta \in \mathbb{R}$. La familia de distribuciones normales $\mathcal{N}(\theta, \sigma^2)$, $\theta \in \mathbb{R}$, es una familia regular uniparamétrica con densidades de la forma

$$f(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}.$$

Usando el resultado del Lema 3.5 se puede ver que *el estimador de máxima verosimilitud para θ es*

$$\hat{\theta}_{mv}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

En efecto, como

$$\psi(\theta|x) = \frac{\partial \log f(x|\theta)}{\partial \theta} = \frac{x - \theta}{\sigma^2}$$

la ecuación de verosimilitud (24) equivale a

$$\sum_{i=1}^n (x_i - \theta) = 0.$$

El resultado se obtiene despejando θ . □

Ejemplo 3.9 (Distribuciones normales). La familia de distribuciones normales

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

es una familia regular con parámetro bidimensional $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. Para encontrar el estimador de máxima verosimilitud del parámetro (μ, σ^2) basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ usaremos los resultados del Lema 3.5. La densidad de cada variable X es

$$f(x|\mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

con lo cual

$$\log f(x|\mu, \sigma^2) = \log(2\pi)^{-\frac{1}{2}} - \frac{1}{2} \log \sigma^2 - \frac{(x - \mu)^2}{2\sigma^2}.$$

En consecuencia,

$$\frac{\partial \log f(x|\mu, \sigma^2)}{\partial \mu} = \frac{x - \mu}{\sigma^2}$$

y

$$\frac{\partial \log f(x|\mu, \sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(x - \mu)^2}{2(\sigma^2)^2}.$$

Luego el sistema de ecuaciones (22) se transforma en el sistema

$$\begin{aligned} \frac{1}{\sigma^2} \left(\sum_{i=1}^n x_i - n\mu \right) &= 0, \\ \frac{1}{2\sigma^2} \left(-n + \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) &= 0. \end{aligned}$$

que tiene como solución

$$\begin{aligned} \mu &= \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \\ \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned}$$

Se puede comprobar que en ese punto de coordenadas (μ, σ^2) se alcanza el máximo absoluto de la función $\log L(\mu, \sigma^2 | \mathbf{x})$.

Resumiendo, cuando la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ arroja los valores $\mathbf{x} = (x_1, \dots, x_n)$, el estimador de máxima verosimilitud para (μ, σ^2) es el punto del conjunto paramétrico $\mathbb{R} \times (0, \infty)$ cuyas coordenadas son el promedio y la varianza muestrales: $\hat{\mu}_{mv}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$ y $\hat{\sigma}^2_{mv}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

Por lo tanto, el *estimador de máxima verosimilitud* para (μ, σ^2) , basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de variables normales, $\mathcal{N}(\mu, \sigma^2)$, es el punto en $\mathbb{R} \times (0, \infty)$ de coordenadas aleatorias

$$\hat{\mu}_{mv}(\mathbf{X}) = \bar{X}, \quad \hat{\sigma}^2_{mv}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (30)$$

□

3.2.1. Familias exponenciales

Muchos modelos estadísticos pueden considerarse como casos particulares de una familia más general de distribuciones.

Definición 3.10 (Familias exponenciales). Decimos que la distribución de una variable aleatoria X pertenece a una *familia exponencial unidimensional* de distribuciones, si podemos escribir su función de probabilidad o su función densidad como

$$f(x|\theta) = e^{a(\theta)T(x)+b(\theta)+S(x)}, \quad x \in \mathbb{S}, \quad (31)$$

donde, a y b son funciones de θ ; T y S son funciones de x y \mathbb{S} no depende de θ .

Nota Bene. Si las funciones a y b son derivables y el espacio paramétrico Θ es abierto, las densidades (31) constituyen una familia regular uniparamétrica y en consecuencia, para encontrar el estimador de máxima verosimilitud de θ , basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$, se puede usar el resultado del Lema 3.5.

Debido a que el logaritmo de la densidad (31) es

$$\log f(x|\theta) = a(\theta)T(x) + b(\theta) + S(x)$$

tenemos que

$$\psi(\theta|x) = \frac{\partial \log f(x|\theta)}{\partial \theta} = a'(\theta)T(x) + b'(\theta)$$

y en consecuencia, la ecuación de verosimilitud (24) adopta la forma

$$a'(\theta) \sum_{i=1}^n T(x_i) + nb'(\theta) = 0.$$

Por lo tanto, el estimador de máxima verosimilitud para θ satisface la ecuación

$$\frac{-b'(\theta)}{a'(\theta)} = \frac{1}{n} \sum_{i=1}^n T(x_i). \quad (32)$$

Ejemplo 3.11 (Distribuciones exponenciales). Sea X una variable aleatoria con distribución Exponencial(λ), $\lambda > 0$. Podemos escribir

$$f(x|\lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \log \lambda}$$

Por lo tanto, la distribución de X pertenece a una familia exponencial unidimensional con $a(\lambda) = -\lambda$, $b(\lambda) = \log \lambda$, $T(x) = x$, $S(x) = 0$ y $\mathbb{S} = (0, \infty)$. La ecuación de verosimilitud (32) adopta la forma

$$\frac{1}{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \quad (33)$$

cuya solución es $\lambda = 1/\bar{x}$. Se puede verificar que el valor de λ así obtenido maximiza el logaritmo de la verosimilitud.

Si la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ arrojó los valores $\mathbf{x} = (x_1, \dots, x_n)$, el estimador de máxima verosimilitud para λ es

$$\hat{\lambda}_{mv}(\mathbf{x}) = (\bar{x})^{-1}.$$

Por lo tanto, el *estimador de máxima verosimilitud* para λ , basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de variables con distribución Exponencial(λ), es

$$\hat{\lambda}_{mv}(\mathbf{X}) = \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^{-1}.$$

□

Ejemplo 3.12 (Distribuciones normales con media conocida). Sea X una variable aleatoria con distribución normal $\mathcal{N}(\mu, \sigma^2)$, donde la media μ es conocida y la varianza $\sigma^2 > 0$. Podemos escribir

$$f(x|\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = e^{-\frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2} \log \sigma^2 - \log \sqrt{2\pi}}$$

Por lo tanto, la distribución de X pertenece a una familia exponencial unidimensional con $a(\sigma^2) = -\frac{1}{2\sigma^2}$, $b(\sigma^2) = -\frac{1}{2} \log \sigma^2$, $T(x) = (x - \mu)^2$, $S(x) = -\log \sqrt{2\pi}$ y $\mathbb{S} = \mathbb{R}$. La ecuación de verosimilitud (32) adopta la forma

$$\frac{1/2\sigma^2}{1/2(\sigma^2)^2} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (34)$$

cuya solución es $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$. Se puede verificar que el valor de σ^2 así obtenido maximiza el logaritmo de la verosimilitud.

Si la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ arrojó los valores $\mathbf{x} = (x_1, \dots, x_n)$, el estimador de máxima verosimilitud para σ^2 es

$$\widehat{\sigma^2}_{mv}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2.$$

Por lo tanto, el *estimador de máxima verosimilitud* para σ^2 , basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de variables con distribución $\mathcal{N}(\mu, \sigma^2)$, es

$$\widehat{\sigma^2}_{mv}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

□

3.2.2. Malas noticias!

“*Esta calle es más angosta de lo que pensás*”.
(Proverbio Zen)

Ejemplo 3.13 (Fiabilidad). Sea T_1, \dots, T_n una muestra aleatoria del tiempo de duración sin fallas de una máquina cuya función intensidad de fallas es $\lambda(t) = \beta t^{\beta-1} \mathbf{1}\{t > 0\}$, donde el parámetro de “desgaste” $\beta > 0$ es desconocido. La densidad de cada tiempo T es

$$f(t|\beta) = \beta t^{\beta-1} e^{-t^\beta} \mathbf{1}\{t > 0\} \quad (35)$$

Observando que

$$\log f(t|\beta) = \log \beta + (\beta - 1) \log t - t^\beta$$

y derivando respecto de β se obtiene

$$\frac{\partial \log f(x|\beta)}{\partial \beta} = \frac{1}{\beta} + \log t - t^\beta \log t.$$

Por lo tanto, la ecuación de verosimilitud (24) adopta la forma

$$\frac{n}{\beta} + \sum_{i=1}^n \log t_i - \sum_{i=1}^n t_i^\beta \log t_i = 0 \quad (36)$$

La mala noticia es que la ecuación (36) no tiene una solución analítica explícita. \square

El ejemplo anterior muestra que en algunos casos la ecuación de verosimilitud no presenta solución analítica explícita. En tales casos, los estimadores de máxima verosimilitud pueden obtenerse mediante métodos numéricos.

Método de Newton-Raphson. El método de Newton-Raphson es un procedimiento iterativo para obtener una raíz de una ecuación

$$g(\theta) = 0, \quad (37)$$

donde $g(\cdot)$ es una función suave. La idea es la siguiente: supongamos que θ es una raíz de la ecuación (37). Desarrollando $g(\cdot)$ en serie de Taylor en torno de un punto θ_0 , obtenemos que

$$g(\theta) \approx g(\theta_0) + (\theta - \theta_0)g'(\theta_0).$$

En consecuencia, si θ_0 está cerca de una raíz θ de la ecuación (37), debería ocurrir lo siguiente

$$\theta \approx \theta_0 - \frac{g(\theta_0)}{g'(\theta_0)}. \quad (38)$$

De la ecuación (38) obtenemos el procedimiento iterativo

$$\theta_{j+1} = \theta_j - \frac{g(\theta_j)}{g'(\theta_j)} \quad (39)$$

que se inicia con un valor θ_0 y produce un nuevo valor θ_1 a partir de (39) y así siguiendo, hasta que el proceso se estabilice, o sea, hasta que $|\theta_{j+1} - \theta_j| < \epsilon$ para un $\epsilon > 0$ “pequeño” y prefijado.

Ejemplo 3.14 (Continuación del Ejemplo 3.13). Para resolver la ecuación (36) usaremos el procedimiento de Newton-Raphson aplicado a la función

$$g(\beta) = \frac{n}{\beta} + \sum_{i=1}^n \log t_i - \sum_{i=1}^n t_i^\beta \log t_i.$$

Como

$$g'(\beta) = -\frac{n}{\beta^2} - \sum_{i=1}^n t_i^\beta (\log t_i)^2,$$

el procedimiento iterativo (39) adopta la forma

$$\beta_{j+1} = \beta_j + \frac{\frac{n}{\beta} + \sum_{i=1}^n \log t_i - \sum_{i=1}^n t_i^\beta \log t_i}{\frac{n}{\beta^2} + \sum_{i=1}^n t_i^\beta (\log t_i)^2}. \quad (40)$$

Generando una muestra aleatoria de tamaño $n = 20$ de una variable aleatoria T con densidad dada por (35) con $\beta = 2$ e inicializando el procedimiento iterativo (40) con $\beta_1 = \bar{T}$ obtuvimos que $\hat{\beta}_{mv} = 2.3674$.

Generando una muestra aleatoria de tamaño $n = 10000$ de una variable aleatoria T con densidad dada por (35) con $\beta = 2$ e inicializando el procedimiento iterativo (40) con $\beta_1 = \bar{T}$ obtuvimos que $\hat{\beta}_{mv} = 1.9969$. \square

3.3. Cálculo del emv para familias no regulares

Venía rápido, muy rápido y se le soltó un patín ...

Ahora mostraremos algunos ejemplos correspondientes a familias no regulares. En estos casos hay que analizar dónde se realiza el máximo “a mano”.

Ejemplo 3.15 (Distribuciones de Bernoulli con parámetros discretos). Supongamos que los valores observados en la secuencia (29) que aparece en el Ejemplo 3.7 fueron arrojados por una muestra aleatoria de tamaño $n = 20$ de una variable aleatoria X con distribución Bernoulli(p), donde $p = 0.45$ o $p = 0.65$. La familia de distribuciones no es regular debido a que el espacio paramétrico $\{0.45, 0.65\}$ no es abierto. En esta situación no puede utilizarse la metodología del Lema 3.5 pues conduce a resultados totalmente disparatados. Lo único que se puede hacer es comparar los valores $L(0.45|\mathbf{x})$, $L(0.65|\mathbf{x})$ y quedarse con el valor de $p \in \{0.45, 0.65\}$ que haga máxima la probabilidad de observar el resultado \mathbf{x} :

$$\begin{aligned} L(0.45|\mathbf{x}) &= (0.45)^{11}(0.55)^9 = (7.0567...)10^{-7} \\ L(0.65|\mathbf{x}) &= (0.65)^{11}(0.35)^9 = (6.8969...)10^{-7}. \end{aligned}$$

Por lo tanto, el estimador de máxima verosimilitud, basado en las observaciones (29), será

$$\hat{p}_{mv}(0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0) = 0.45.$$

\square

Ejemplo 3.16 (Distribución uniforme). La familia $\{\mathcal{U}(0, \theta) : \theta > 0\}$ de distribuciones uniformes no es una familia regular debido a que el soporte de la densidad de la distribución $\mathcal{U}(0, \theta)$ es $[0, \theta]$ (y depende claramente del valor del parámetro θ). En esta situación tampoco puede utilizarse la metodología del Lema 3.5. En este caso $\Theta = (0, \infty)$ y las funciones de densidad son de la forma

$$f(x|\theta) = \frac{1}{\theta} \mathbf{1}\{0 \leq x \leq \theta\}.$$

La función de verosimilitud es

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^n \frac{1}{\theta} \mathbf{1}\{0 \leq x_i \leq \theta\} = \frac{1}{\theta^n} \prod_{i=1}^n \mathbf{1}\{0 \leq x_i \leq \theta\} \\ &= \frac{1}{\theta^n} \mathbf{1}\left\{\max_{i=1,\dots,n} x_i \leq \theta\right\}. \end{aligned}$$

Si $\theta < \max_i x_i$, entonces $L(\theta|\mathbf{x}) = 0$. Si $\theta \geq \max_i x_i$, entonces $L(\theta|\mathbf{x}) = \theta^{-n}$, una función decreciente en θ . En consecuencia, su máximo se alcanza en

$$\theta = \max_{i=1,\dots,n} x_i.$$

Por lo tanto, el estimador de máxima verosimilitud para θ , basado en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de una variable aleatoria $X \sim \mathcal{U}(0, \theta)$, es el máximo de la muestra

$$\hat{\theta}_{mv}(\mathbf{X}) = X_{(n)} := \max_{i=1,\dots,n} X_i.$$

□

Ejemplo 3.17 (Distribución uniforme). La familia $\{\mathcal{U}(\theta - 1/2, \theta + 1/2) : \theta \in \mathbb{R}\}$ de distribuciones uniformes no es una familia regular debido a que el soporte de la densidad de la distribución $\mathcal{U}(\theta - 1/2, \theta + 1/2)$ es $[\theta - 1/2, \theta + 1/2]$ (y depende claramente del valor del parámetro θ). En este caso $\Theta = \mathbb{R}$ y las funciones de densidad son de la forma

$$f(x|\theta) = \mathbf{1}\{\theta - 1/2 \leq x \leq \theta + 1/2\}.$$

La función de verosimilitud es

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^n \mathbf{1}\{\theta - 1/2 \leq x_i \leq \theta + 1/2\} \\ &= \mathbf{1}\left\{\max_{i=1,\dots,n} x_i - 1/2 \leq \theta \leq \min_{i=1,\dots,n} x_i + 1/2\right\} \\ &= \mathbf{1}\{x_{(n)} - 1/2 \leq \theta \leq x_{(1)} + 1/2\}, \end{aligned}$$

pues

$$\theta - 1/2 \leq x_i \leq \theta + 1/2, \quad i = 1, \dots, n,$$

si y solamente si

$$\theta \leq x_i + 1/2 \quad \text{y} \quad x_i - 1/2 \leq \theta, \quad i = 1, \dots, n,$$

Como $L(\theta|\mathbf{x})$ se anula para $\theta < x_{(n)}$ y para $\theta > x_{(1)} + 1/2$ y es constantemente 1 en el intervalo $[x_{(n)} - 1/2, x_{(1)} + 1/2]$, tenemos que cualquier punto de ese intervalo es un estimador de máxima verosimilitud para θ . En particular,

$$\hat{\theta}(\mathbf{x}) = \frac{x_{(1)} + x_{(n)}}{2}$$

es un estimador de máxima verosimilitud para θ . Etc... □

3.4. Principio de invariancia

En lo que sigue presentamos una propiedad bastante importante del método de máxima verosimilitud.

Teorema 3.18 (Principio de invariancia). *Sea X_1, \dots, X_n una muestra aleatoria de una variable aleatoria X cuya distribución pertenece a la familia paramétrica $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$. Sea $g : \Theta \rightarrow \Lambda$ una función biunívoca de Θ sobre Λ . Si $\hat{\theta}$ es un estimador de máxima verosimilitud para θ , entonces $g(\hat{\theta})$ es un estimador de máxima verosimilitud para $\lambda = g(\theta)$.*

Demostración. Como $\lambda = g(\theta)$ es una función biunívoca de Θ sobre Λ , la función de verosimilitud $L(\theta|\mathbf{x})$ se puede expresar en función de λ ya que $\theta = g^{-1}(\lambda)$. Denominemos a la función de verosimilitud, como función de λ , por $L^*(\lambda|\mathbf{x})$. Es claro que

$$L^*(\lambda|\mathbf{x}) = L(g^{-1}(\lambda)|\mathbf{x}).$$

Sea $\hat{\theta}_{mv} \in \Theta$ un estimador de máxima verosimilitud para θ y sea $\hat{\lambda} := g(\hat{\theta}_{mv}) \in \Lambda$ su imagen por g . Hay que mostrar que vale lo siguiente:

$$L^*(\hat{\lambda}|\mathbf{x}) = \max_{\lambda \in \Lambda} L^*(\lambda|\mathbf{x})$$

Pero esto es inmediato, debido a que

$$\begin{aligned} L^*(\hat{\lambda}|\mathbf{x}) &= L(g^{-1}(\hat{\lambda})|\mathbf{x}) = L(\hat{\theta}_{mv}|\mathbf{x}) = \max_{\theta \in \Theta} L(\theta|\mathbf{x}) = \max_{\lambda \in \Lambda} L(g^{-1}(\lambda)|\mathbf{x}) \\ &= \max_{\lambda \in \Lambda} L^*(\lambda|\mathbf{x}). \end{aligned}$$

Por lo tanto,

$$\widehat{g(\theta)}_{mv} = g(\hat{\theta}_{mv}).$$

□

Ejemplo 3.19. Sea X_1, \dots, X_n una muestra aleatoria de la variable aleatoria $X \sim \mathcal{N}(\mu, 1)$. En el Ejemplo 3.8 vimos que $\hat{\mu}_{mv} = \bar{X}$ es el estimador de máxima verosimilitud para μ . Queremos estimar

$$g(\mu) = \mathbb{P}_\mu(X \leq 0) = \Phi(-\mu).$$

Por el principio de invariancia, tenemos que

$$g(\hat{\mu}_{mv}) = \Phi(-\bar{X})$$

es el estimador de máxima verosimilitud para $\mathbb{P}_\mu(X \leq 0)$.

□

Nota Bene En general, si $\lambda = g(\theta)$, aunque g no sea biunívoca, se define el estimador de máxima verosimilitud de λ por

$$\hat{\lambda} = g(\hat{\theta}_{mv}).$$

4. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Bolfarine, H., Sandoval, M. C.: Introdução à Inferência Estatística. SBM, Rio de Janeiro. (2001).
2. Borovkov, A. A.: Estadística matemática. Mir, Moscú. (1984).
3. Cramer, H.: Métodos matemáticos de estadística. Aguilar, Madrid. (1970).
4. Hoel P. G.: Introducción a la estadística matemática. Ariel, Barcelona. (1980).
5. Maronna R.: Probabilidad y Estadística Elementales para Estudiantes de Ciencias. Editorial Exacta, La Plata. (1995).

Estimación por intervalo
(Borradores, Curso 23)

Sebastian Grynberg

27-29 de mayo de 2013



Si ves al futuro, dile que no venga.
(Juan José Castelli)

Índice

1. Estimación por intervalo	3
1.1. El método del pivote	5
1.1.1. Pivotes decrecientes	5
1.1.2. Pivotes crecientes	8
2. Muestras de Poblaciones Normales	10
2.1. Media y varianza desconocidas	10
2.1.1. Teorema llave	10
2.1.2. Cotas e intervalos de confianza para la varianza	11
2.1.3. Cotas e intervalos de confianza para la media	12
2.1.4. Ejemplo	13
2.2. Media de la normal con varianza conocida	13
2.3. Varianza de la normal con media conocida	14
3. Intervalos aproximados para ensayos Bernoulli	15
4. Comparación de dos muestras normales	17
4.1. Cotas e intervalos de confianza para la diferencia de medias	17
4.1.1. Varianzas conocidas	17
4.1.2. Varianzas desconocidas.	17
4.2. Cotas e intervalos de confianza para el cociente de varianzas.	19
5. Comparación de dos muestras	19
5.1. Planteo general	19
5.2. Problema de dos muestras binomiales	20
6. Apéndice: Demostración del Teorema llave	22
6.1. Preliminares de Análisis y Álgebra	22
6.2. Lema previo	23
6.3. Demostración del Teorema.	23
7. Bibliografía consultada	24

1. Estimación por intervalo

En lo que sigue consideramos el problema de estimación de parámetros utilizando intervalos de confianza. Consideramos una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de la variable aleatoria X cuya función de distribución $F(x) := \mathbb{P}(X \leq x)$, pertenece a la familia paramétrica de distribuciones (distinguibles) $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$. La idea básica es la siguiente: aunque no podamos determinar exactamente el valor de θ podemos tratar de construir un intervalo aleatorio $[\theta^-, \theta^+]$ tal que con una probabilidad bastante alta, sea capaz de “capturar” el valor desconocido θ .

Definición 1.1 (Intervalo de confianza). Un *intervalo de confianza* para θ de nivel β es un intervalo aleatorio, $I(\mathbf{X})$, que depende de la muestra aleatoria \mathbf{X} , tal que

$$\mathbb{P}_\theta(\theta \in I(\mathbf{X})) = \beta, \quad (1)$$

para todo $\theta \in \Theta$.

Definición 1.2 (Cotas de confianza). Una *cota inferior de confianza* para θ , de nivel β , basada en la muestra aleatoria \mathbf{X} , es una variable aleatoria $\theta_1(\mathbf{X})$ tal que

$$\mathbb{P}_\theta(\theta_1(\mathbf{X}) \leq \theta) = \beta, \quad (2)$$

para todo $\theta \in \Theta$.

Una *cota superior de confianza* para θ , de nivel β , basada en la muestra aleatoria \mathbf{X} , es una variable aleatoria $\theta_2(\mathbf{X})$ tal que

$$\mathbb{P}_\theta(\theta \leq \theta_2(\mathbf{X})) = \beta, \quad (3)$$

para todo $\theta \in \Theta$.

Nota Bene. En el caso discreto no siempre se pueden obtener las igualdades (1), (2) o (3). Para evitar este tipo de problemas se suele definir un intervalo mediante la condición más laxa $\mathbb{P}_\theta(\theta \in I(\mathbf{X})) \geq \beta, \forall \theta$. En este caso el $\min_\theta P_\theta(\theta \in I(\mathbf{X}))$ se llama *nivel de confianza*.

Observación 1.3. Sean $\theta_1(\mathbf{X})$ una cota inferior de confianza de nivel $\beta_1 > 1/2$ y $\theta_2(\mathbf{X})$ una cota superior de confianza de nivel $\beta_2 > 1/2$, tales que $\mathbb{P}_\theta(\theta_1(\mathbf{X}) \leq \theta_2(\mathbf{X})) = 1$ para todo $\theta \in \Theta$. Entonces,

$$I(\mathbf{X}) = [\theta_1(\mathbf{X}), \theta_2(\mathbf{X})]$$

define un intervalo de confianza para θ de nivel $\beta = \beta_1 + \beta_2 - 1$. En efecto,

$$\begin{aligned} \mathbb{P}_\theta(\theta \in I(\mathbf{X})) &= 1 - \mathbb{P}_\theta(\theta < \theta_1(\mathbf{X}) \text{ o } \theta > \theta_2(\mathbf{X})) \\ &= 1 - \mathbb{P}_\theta(\theta < \theta_1(\mathbf{X})) - \mathbb{P}_\theta(\theta > \theta_2(\mathbf{X})) \\ &= 1 - (1 - \beta_1) - (1 - \beta_2) = \beta_1 + \beta_2 - 1. \end{aligned} \quad (4)$$

La identidad (4) muestra que la construcción de intervalos de confianza se reduce a la construcción de cotas inferiores y superiores. Más precisamente, *si se quiere construir un intervalo de confianza de nivel β , basta construir una cota inferior de nivel $\beta_1 = (1 + \beta)/2$ y una cota superior de nivel $\beta_2 = (1 + \beta)/2$* . \square

Las ideas principales para construir intervalos de confianza están contenidas en el ejemplo siguiente.

Ejemplo 1.4 (Media de la normal con varianza conocida). Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria $X \sim \mathcal{N}(\mu, \sigma^2)$, con varianza σ^2 conocida. Para obtener un intervalo de confianza de nivel β para μ , consideramos el estimador de máxima verosimilitud para μ

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

La distribución de \bar{X} se obtiene utilizando los resultados conocidos sobre sumas de normales independientes y de cambio de escala:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

En consecuencia,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1).$$

Por lo tanto, para cada $\mu \in \mathbb{R}$ vale que

$$\mathbb{P}_\mu \left(-z_{(1+\beta)/2} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq z_{(1+\beta)/2} \right) = \beta.$$

Despejando μ de las desigualdades dentro de la probabilidad, resulta que

$$\mathbb{P}_\mu \left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{(1+\beta)/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{(1+\beta)/2} \right) = \beta,$$

y por lo tanto el intervalo

$$I(\mathbf{X}) = \left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{(1+\beta)/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{(1+\beta)/2} \right]$$

es un intervalo de confianza para μ de nivel β . □

Nota Bene. Las ideas principales para construir el intervalo de confianza contenidas en el ejemplo anterior son las siguientes:

1. Obtener un estimador del parámetro y caracterizar su distribución.
2. Transformar el estimador de parámetro hasta convertirlo en una variable aleatoria cuya distribución “conocida” que no dependa del parámetro.
3. Poner cotas para el estimador transformado y despejar el parámetro.

1.1. El método del pivote

Cuando se quieren construir intervalos de confianza para θ lo más natural es comenzar la construcción apoyándose en algún estimador puntual del parámetro $\hat{\theta}(\mathbf{X})$ (cuya distribución depende de θ). Una técnica general para construir intervalos de confianza, llamada el *método del pivote*, consiste en transformar el estimador $\hat{\theta}(\mathbf{X})$ hasta convertirlo en una variable aleatoria cuya distribución sea “conocida” y no dependa de θ . Para que la transformación sea útil no debe depender de ningún otro parámetro desconocido.

Definición 1.5 (Pivote). Una variable aleatoria de la forma $Q(\mathbf{X}, \theta)$ se dice una *cantidad pivotal* o un *pivote* para el parámetro θ si su distribución no depende de θ (ni de ningún parámetro desconocido, cuando hay varios parámetros).

Nota Bene. Por definición, la distribución del pivote $Q(\mathbf{X}, \theta)$ no depende de θ . Para cada $\alpha \in (0, 1)$ notaremos mediante q_α el cuantil- α del pivote. Si el pivote tiene distribución continua y su función de distribución es estrictamente creciente, q_α es la única solución de la ecuación

$$\mathbb{P}_\theta(Q(\mathbf{X}, \theta) \leq q_\alpha) = \alpha.$$

Método. Si se consigue construir un pivote $Q(\mathbf{X}, \theta)$ para el parámetro θ , el problema de la construcción de intervalos de confianza, de nivel β , se descompone en dos partes:

1. Encontrar parejas de números reales $a < b$ tales que $\mathbb{P}_\theta(a \leq Q(\mathbf{X}; \theta) \leq b) = \beta$. Por ejemplo, $a = q_{\frac{1-\beta}{2}}$ y $b = q_{\frac{1+\beta}{2}}$.
2. Despejar el parámetro θ de las desigualdades $a \leq Q(\mathbf{X}, \theta) \leq b$.

Si el pivote $Q(\mathbf{X}, \theta)$ es una función monótona en θ se puede ver que existen $\theta_1(\mathbf{X})$ y $\theta_2(\mathbf{X})$ tales que

$$a \leq Q(\mathbf{X}; \theta) \leq b \Leftrightarrow \theta_1(\mathbf{X}) \leq \theta \leq \theta_2(\mathbf{X})$$

y entonces

$$\mathbb{P}_\theta(\theta_1(\mathbf{X}) \leq \theta \leq \theta_2(\mathbf{X})) = \beta,$$

de modo que $I(\mathbf{X}) = [\theta_1(\mathbf{X}), \theta_2(\mathbf{X})]$ es un intervalo de confianza para θ de nivel β . \square

1.1.1. Pivotes decrecientes

Sea $Q(\mathbf{X}, \theta)$ un pivote para θ que goza de las siguientes propiedades:

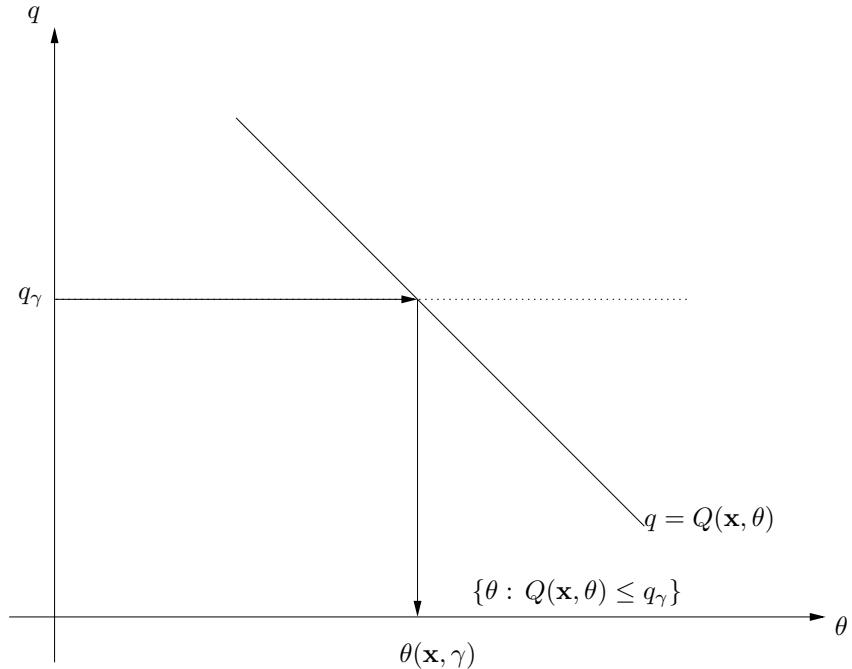
- (i) la función de distribución de $Q(\mathbf{X}, \theta)$ es continua y estrictamente creciente;
- (ii) para cada \mathbf{x} , la función $Q(\mathbf{x}, \theta)$ es continua y monótona decreciente en la variable θ :

$$\theta_1 < \theta_2 \implies Q(\mathbf{x}, \theta_1) > Q(\mathbf{x}, \theta_2)$$

Sea $\gamma \in (0, 1)$, arbitrario pero fijo y sea q_γ el cuantil- γ del pivote $Q(\mathbf{X}, \theta)$.

Para cada \mathbf{x} , sea $\theta(\mathbf{x}, \gamma)$ la única solución de la ecuación en θ

$$Q(\mathbf{x}, \theta) = q_\gamma.$$



Como el pivote $Q(\mathbf{X}, \theta)$ es decreciente en θ tenemos que

$$Q(\mathbf{X}, \theta) \leq q_\gamma \iff \theta(\mathbf{X}, \gamma) \leq \theta.$$

En consecuencia,

$$\mathbb{P}_\theta(\theta(\mathbf{X}, \gamma) \leq \theta) = \mathbb{P}_\theta(Q(\mathbf{X}, \theta) \leq q_\gamma) = \gamma, \quad \forall \theta \in \Theta.$$

Por lo tanto, $\theta(\mathbf{X}, \gamma)$ es una cota inferior de confianza para θ de nivel γ y una cota superior de nivel $1 - \gamma$.

Método

Sea $\beta \in (0, 1)$. Si se dispone de un pivote $Q(\mathbf{X}, \theta)$ que satisface las propiedades (i) y (ii) enunciadas más arriba, entonces

- la variable aleatoria, $\theta_1(\mathbf{X})$, que se obtiene resolviendo la ecuación $Q(\mathbf{X}, \theta) = q_\beta$ es una *cota inferior de confianza* para θ , de nivel β .
- la variable aleatoria, $\theta_2(\mathbf{X})$, que se obtiene resolviendo la ecuación $Q(\mathbf{X}, \theta) = q_{1-\beta}$ es una *cota superior de confianza* para θ , de nivel β .
- el intervalo aleatorio $I(\mathbf{X}) = [\theta_1(\mathbf{X}), \theta_2(\mathbf{X})]$ cuyos extremos son las soluciones respectivas de las ecuaciones $Q(\mathbf{X}, \theta) = q_{\frac{1+\beta}{2}}$ y $Q(\mathbf{X}, \theta) = q_{\frac{1-\beta}{2}}$, es un *intervalo “bilateral” de confianza* para θ , de nivel β .

Ejemplo 1.6 (Extremo superior de la distribución uniforme). Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria $X \sim \mathcal{U}(0, \theta)$, $\theta > 0$.

El estimador de máxima verosimilitud para θ es $X_{(n)} = \max(X_1, \dots, X_n)$ y tiene densidad de la forma

$$f(x) = \frac{nx^{n-1}}{\theta^n} \mathbf{1}\{0 \leq x \leq \theta\}.$$

Como la distribución de $X_{(n)}$ depende de θ , $X_{(n)}$ no es un pivote para θ . Sin embargo, podemos liberarnos de θ utilizando un cambio de variables lineal de la forma $Q = X_{(n)}/\theta$:

$$f_Q(q) = nq^{n-1} \mathbf{1}\{0 \leq q \leq 1\}.$$

Por lo tanto,

$$Q(\mathbf{X}, \theta) = X_{(n)}/\theta$$

es un pivote para θ .

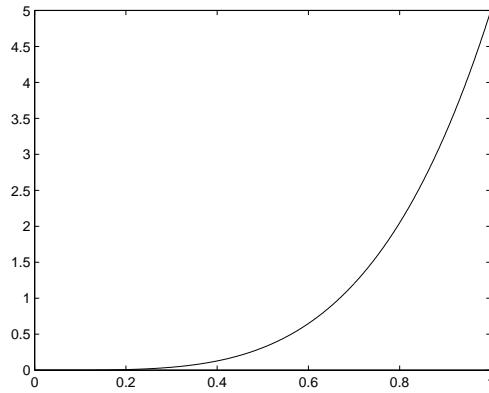


Figura 1: Forma típica del gráfico de la densidad del pivote $Q(\mathbf{X}, \theta)$.

Los cuantiles- γ para Q se obtienen observando que

$$\gamma = \mathbb{P}(Q(\mathbf{X}, \theta) \leq q_\gamma) = \int_0^{q_\gamma} f_Q(q) dq \iff q_\gamma = \gamma^{1/n}.$$

Construyendo un intervalo de confianza. Dado el nivel de confianza $\beta \in (0, 1)$, para construir un intervalo de confianza de nivel β notamos que

$$\beta = \mathbb{P}_\theta(q_{1-\beta} \leq Q(\mathbf{X}, \theta) \leq 1) = \mathbb{P}_\theta(q_{1-\beta} \leq X_{(n)}/\theta \leq 1)$$

Despejando θ de las desigualdades dentro de la probabilidad, resulta que

$$I(\mathbf{X}) = \left[X_{(n)}, \frac{X_{(n)}}{q_{1-\beta}} \right] = \left[X_{(n)}, \frac{X_{(n)}}{(1-\beta)^{1/n}} \right]$$

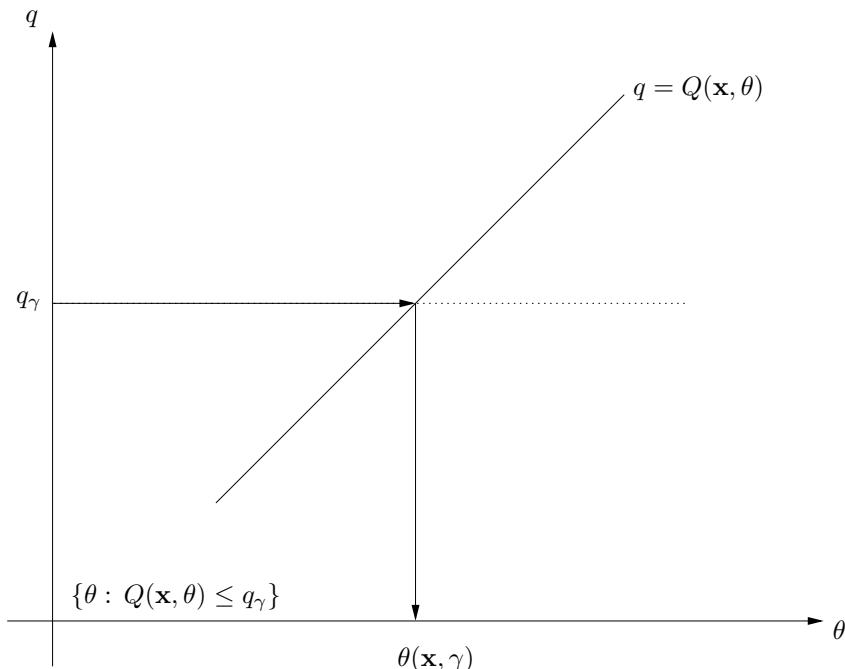
es un intervalo de confianza para θ de nivel β . □

1.1.2. Pivotes crecientes

Sea $Q(\mathbf{X}, \theta)$ un pivote para θ que goza de las siguientes propiedades:

- (i) la función de distribución de $Q(\mathbf{X}, \theta)$ es continua y estrictamente creciente;
- (ii') para cada \mathbf{x} , la función $Q(\mathbf{x}, \theta)$ es continua y monótona creciente en la variable θ :

$$\theta_1 < \theta_2 \implies Q(\mathbf{x}, \theta_1) < Q(\mathbf{x}, \theta_2)$$



Sea $\gamma \in (0, 1)$, arbitrario pero fijo y sea q_γ el cuantil- γ del pivote $Q(\mathbf{X}, \theta)$. Para cada \mathbf{x} , sea $\theta(\mathbf{x}, \gamma)$ la única solución de la ecuación en θ

$$Q(\mathbf{x}, \theta) = q_\gamma.$$

Como el pivote $Q(\mathbf{X}, \theta)$ es creciente en θ tenemos que

$$Q(\mathbf{X}, \theta) \leq q_\gamma \iff \theta \leq \theta(\mathbf{X}, \gamma).$$

En consecuencia,

$$\mathbb{P}_\theta(\theta \leq \theta(\mathbf{X}, \gamma)) = \mathbb{P}_\theta(Q(\mathbf{X}, \theta) \leq q_\gamma) = \gamma, \quad \forall \theta \in \Theta.$$

Por lo tanto, $\theta(\mathbf{X}, \gamma)$ es una cota superior de confianza para θ de nivel γ y una cota inferior de nivel $1 - \gamma$.

Método

Sea $\beta \in (0, 1)$. Si se dispone de un pivote $Q(\mathbf{X}, \theta)$ que satisface las propiedades (i) y (ii') enunciadas más arriba, entonces

- la variable aleatoria, $\theta_1(\mathbf{X})$, que se obtiene resolviendo la ecuación $Q(\mathbf{X}, \theta) = q_{1-\beta}$ es una *cota inferior de confianza* para θ , de nivel β .
- la variable aleatoria, $\theta_2(\mathbf{X})$, que se obtiene resolviendo la ecuación $Q(\mathbf{X}, \theta) = q_\beta$ es una *cota superior de confianza* para θ , de nivel β .
- el intervalo aleatorio $I(\mathbf{X}) = [\theta_1(\mathbf{X}), \theta_2(\mathbf{X})]$, cuyos extremos son las soluciones respectivas de las ecuaciones $Q(\mathbf{X}, \theta) = q_{\frac{1-\beta}{2}}$ y $Q(\mathbf{X}, \theta) = q_{\frac{1+\beta}{2}}$, es un *intervalo “bilateral” de confianza* para θ , de nivel β .

Ejemplo 1.7 (Intensidad de la distribución exponencial). Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria $X \sim \text{Exp}(\lambda)$, $\lambda > 0$.

El estimador de máxima verosimilitud para λ es $1/\bar{X}$, donde $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Sabemos que la suma $n\bar{X} = \sum_{i=1}^n X_i$ tiene distribución $\Gamma(n, \lambda)$.

Como la distribución de $n\bar{X}$ depende de λ , $n\bar{X}$ no es un pivote para λ . Sin embargo, podemos liberarnos de λ utilizando un cambio de variables lineal de la forma $Q = an\bar{X}$, donde a es positivo y elegido adecuadamente para nuestros propósitos. Si $a > 0$ y $Q = an\bar{X}$, entonces $Q \sim \Gamma(n, \frac{\lambda}{a})$. Poniendo $a = 2\lambda$, resulta que $Q = 2\lambda n\bar{X} \sim \Gamma(n, \frac{1}{2}) = \chi_{2n}^2$. (Recordar que $\Gamma(\frac{n}{2}, \frac{1}{2}) = \chi_n^2$.)

Por lo tanto,

$$Q(\mathbf{X}, \lambda) = 2\lambda n\bar{X} = 2\lambda \sum_{i=1}^n X_i \sim \chi_{2n}^2$$

es un pivote para λ .

Construyendo una cota superior de confianza. Dado $\beta \in (0, 1)$, para construir una cota superior de confianza para λ , de nivel β , primero observamos que el pivote $Q(\mathbf{X}, \lambda) = 2\lambda n\bar{X}$ es una función continua y decreciente en λ . Debido a que

$$2\lambda n\bar{X} = \chi_\beta^2 \iff \lambda = \frac{\chi_\beta^2}{2n\bar{X}}$$

resulta que

$$\lambda_2(\mathbf{X}) = \frac{\chi_\beta^2}{2 \sum_{i=1}^n X_i}$$

es una cota superior de confianza para λ de nivel β .

Ilustración. Consideremos ahora las siguientes 10 observaciones

$$0.5380, 0.4470, 0.2398, 0.5365, 0.0061, \\ 0.3165, 0.0086, 0.0064, 0.1995, 0.9008.$$

En tal caso tenemos $\sum_{i=1}^{10} X_i = 3.1992$. Tomando $\beta = 0.975$, tenemos de la tabla de la distribución χ_{20}^2 que $\chi_{20, 0.975}^2 = 34.17$, entonces $\lambda_2(\mathbf{x}) = 5.34$ es una cota superior de confianza para λ de nivel $\beta = 0.975$. \square

2. Muestras de Poblaciones Normales

En esta sección estudiaremos la distribución de probabilidades de los estimadores de máxima verosimilitud para la media y la varianza de poblaciones normales. La técnica de análisis se basa en la construcción de pivotes para los parámetros desconocidos. Usando esos pivotes mostraremos como construir intervalos de confianza en los distintos escenarios posibles que se pueden presentar.

Notación. En todo lo que sigue usaremos la siguiente notación: para cada $\gamma \in (0, 1)$, z_γ será el único número real tal que $\Phi(z_\gamma) = \gamma$. Gráficamente, a izquierda del punto z_γ el área bajo la campana de Gauss es igual a γ .

Nota Bene. De la simetría de la campana de Gauss, se deduce que para cada $\beta \in (0, 1)$ vale que $z_{(1-\beta)/2} = -z_{(1+\beta)/2}$. Por lo tanto, para $Z \sim \mathcal{N}(0, 1)$ vale que

$$\mathbb{P}(-z_{(1+\beta)/2} \leq Z \leq z_{(1+\beta)/2}) = \Phi(z_{(1+\beta)/2}) - \Phi(-z_{(1+\beta)/2}) = \frac{1+\beta}{2} - \frac{1-\beta}{2} = \beta.$$

2.1. Media y varianza desconocidas

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria $X \sim \mathcal{N}(\mu, \sigma^2)$, con media μ y varianza desconocidas. *Los estimadores de máxima verosimilitud para la media y la varianza, basados en \mathbf{X} , son, respectivamente,*

$$\hat{\mu}_{mv}(\mathbf{X}) = \bar{X}, \quad \widehat{\sigma^2}_{mv}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (5)$$

2.1.1. Teorema llave

Teorema 2.1 (Llave). *Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una distribución $\mathcal{N}(\mu, \sigma^2)$. Valen las siguientes afirmaciones:*

- (a) $Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$ tiene distribución $\mathcal{N}(0, 1)$.
- (b) $U = \frac{n-1}{\sigma^2} S^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$ tiene distribución χ_{n-1}^2 .
- (c) Z y U son variables aleatorias independientes.

Nota Bene. El calificativo de “llave” para el Teorema 2.1 está puesto para destacar que sus resultados son la clave fundamental en la construcción de intervalos de confianza y de reglas de decisión sobre hipótesis estadísticas para distribuciones normales. La prueba de este Teorema puede verse en el Apéndice.

Corolario 2.2 (Pivotes para la media y la varianza). Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una distribución $\mathcal{N}(\mu, \sigma^2)$. Sean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ y $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Vale que

(a)

$$Q(\mathbf{X}, \sigma^2) = \frac{(n-1)}{\sigma^2} S^2 \quad (6)$$

es un pivote para la varianza σ^2 y su distribución es una chi cuadrado con $n - 1$ grados de libertad (en símbolos, $Q(\mathbf{X}, \sigma^2) \sim \chi_{n-1}^2$).

(b)

$$Q(\mathbf{X}, \mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \quad (7)$$

es un pivote para la media μ y su distribución es una t de Student con $n - 1$ grados de libertad (en símbolos, $Q(\mathbf{X}, \mu) \sim t_{n-1}$).

Demostración.

- (a) Inmediato de la afirmación (b) del Teorema 2.1.
- (b) La afirmación (a) del Teorema 2.1 indica que $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0, 1)$. Pero como σ^2 es un parámetro desconocido, la transformación $\sqrt{n}(\bar{X} - \mu)/\sigma$ es inútil por sí sola para construir un pivote. Sin embargo, la afirmación (c) del Teorema 2.1 muestra que este problema se puede resolver reemplazando la desconocida σ^2 por su estimación insesgada S^2 . Concretamente, tenemos que

$$Q(\mathbf{X}, \mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{S} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{S/\sigma} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{U/(n-1)}},$$

donde $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0, 1)$ y $U = \frac{(n-1)}{\sigma^2} S^2 \sim \chi_{n-1}^2$ son variables aleatorias independientes. En consecuencia, $Q(\mathbf{X}, \mu) \sim t_{n-1}$. \square

2.1.2. Cotas e intervalos de confianza para la varianza

Notar que el pivote para la varianza $Q(\mathbf{X}, \sigma^2)$ definido en (6) goza de las propiedades enunciadas en la sección 1.1.1 para pivotes decrecientes:

- la función de distribución de $Q(\mathbf{X}, \sigma^2)$ es continua y estrictamente creciente;
- para cada \mathbf{x} , la función $Q(\mathbf{x}, \sigma^2)$ es continua y monótona decreciente respecto de σ^2 .

En consecuencia, las cotas e intervalos de confianza para la varianza se pueden construir usando el resolviendo la ecuación $Q(\mathbf{X}, \sigma^2) = \chi_{n-1, \gamma}^2$, donde $\text{chi}_{n-1, \gamma}^2$ designa el cuantil- γ de la distribución chi cuadrado con $n - 1$ grados de libertad.

Observando que

$$Q(\mathbf{X}, \sigma^2) = \chi_{n-1, \gamma}^2 \iff \frac{(n-1)S^2}{\sigma^2} = \chi_{n-1, \gamma}^2 \iff \sigma^2 = \frac{(n-1)S^2}{\chi_{n-1, \gamma}^2}, \quad (8)$$

se deduce que, para cada $\beta \in (0, 1)$,

1.

$$\sigma_1^2(\mathbf{X}) = \frac{(n-1)S^2}{\chi_{n-1, \beta}^2}$$

es una cota inferior de confianza de nivel β para σ^2 ;

2.

$$\sigma_2^2(\mathbf{X}) = \frac{(n-1)S^2}{\chi_{n-1, 1-\beta}^2}$$

es una cota superior de confianza de nivel β para σ^2 ;

3.

$$I(\mathbf{X}) = \left[\frac{(n-1)S^2}{\chi_{n-1, (1+\beta)/2}^2}, \frac{(n-1)S^2}{\chi_{n-1, (1-\beta)/2}^2} \right]$$

es un intervalo de confianza de nivel β para σ^2 . □

2.1.3. Cotas e intervalos de confianza para la media

Notar que el pivote para la media $Q(\mathbf{X}, \mu)$ definido en (7) goza de las propiedades enumeradas en la sección 1.1.1 para pivotes decrecientes:

- la función de distribución de $Q(\mathbf{X}, \mu)$ es continua y estrictamente creciente;
- para cada \mathbf{x} , la función $Q(\mathbf{x}, \mu)$ es continua y monótona decreciente respecto de μ .

En consecuencia, las cotas e intervalos de confianza para la varianza se pueden construir usando el resolviendo la ecuación $Q(\mathbf{X}, \mu) = t_{n-1, \gamma}$, donde $t_{n-1, \gamma}$ designa el cuantil- γ de la distribución t de Student con $n - 1$ grados de libertad.

Observando que

$$Q(\mathbf{X}, \mu) = t_{n-1, \gamma} \iff \frac{\sqrt{n}(\bar{X} - \mu)}{S} = t_{n-1, \gamma} \iff \mu = \bar{X} - \frac{S}{\sqrt{n}}t_{n-1, \gamma}, \quad (9)$$

y usando que la densidad de la distribución t_{n-1} es simétrica respecto del origen (i.e., $t_{n-1, 1-\gamma} = -t_{n-1, \gamma}$), tenemos que, para cada $\beta \in (0.5, 1)$,

1.

$$\mu_1(\mathbf{X}) = \bar{X} - \frac{S}{\sqrt{n}}t_{n-1, \beta}$$

es una cota inferior de confianza de nivel β para μ ;

2.

$$\mu_2(\mathbf{X}) = \bar{X} - \frac{S}{\sqrt{n}}t_{n-1, 1-\beta} = \bar{X} + \frac{S}{\sqrt{n}}t_{n-1, \beta}$$

es una cota superior de confianza de nivel β para μ ;

3.

$$I(\mathbf{X}) = \left[\bar{X} - \frac{S}{\sqrt{n}}t_{n-1, (1+\beta)/2}, \bar{X} + \frac{S}{\sqrt{n}}t_{n-1, (1+\beta)/2} \right]$$

es un intervalo de confianza de nivel β para μ . □

2.1.4. Ejemplo

Para fijar ideas vamos a construir intervalos de confianza de nivel $\beta = 0.95$ para la media y la varianza de una variable normal $\mathcal{N}(\mu, \sigma^2)$, basados en una muestra aleatoria de volumen $n = 8$ que arrojó los resultados siguientes: 9, 14, 10, 12, 7, 13, 11, 12.

El problema se resuelve recurriendo a las tablas de las distribuciones χ^2 y t y haciendo algunas cuentas.

Como $n = 8$ consultamos las tablas de χ^2_7 y de t_7 . Para el nivel $\beta = 0.95$ tenemos que $(1+\beta)/2 = 0.975$ y $(1-\beta)/2 = 0.025$. De acuerdo con las tablas $\chi^2_{7,0.975} = 16.0127$, $\chi^2_{7,0.025} = 1.6898$ y $t_{7,0.975} = 2.3646$. Por otra parte, $\bar{X} = 11$, $S^2 = 36/7 = 5.1428$ y $S = 2.2677$.

Algunas cuentas más (y un poco de paciencia) permiten rematar este asunto. Salvo errores de cuentas, $I_1 = [2.248, 21.304]$ es un intervalo de confianza de nivel 0.95 para la varianza, mientras que $I_2 = [9.104, 12.895]$ es un intervalo de confianza de nivel 0.95 para la media. \square

2.2. Media de la normal con varianza conocida

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria $X \sim \mathcal{N}(\mu, \sigma^2)$, con varianza σ^2 conocida. En el Ejemplo 1.4 mostramos que

$$Q(\mathbf{X}, \mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

es un pivote para la media μ .

Como el pivote para la media goza de las propiedades enunciadas en la sección 1.1.1 para pivotes decrecientes,

- la función de distribución de $Q(\mathbf{X}, \mu)$ es continua y estrictamente creciente,
- para cada \mathbf{x} , la función $Q(\mathbf{x}, \mu)$ es continua y monótona decreciente respecto de μ ,

las cotas e intervalos de confianza para la media se pueden construir resolviendo la ecuación $Q(\mathbf{X}, \mu) = z_\gamma$, donde z_γ designa el cuantil- γ de la distribución normal estándar $\mathcal{N}(0, 1)$.

Observando que

$$Q(\mathbf{X}, \mu) = z_\gamma \iff \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} = z_\gamma \iff \mu = \bar{X} - \frac{\sigma}{\sqrt{n}}z_\gamma,$$

y usando que la densidad de la distribución $\mathcal{N}(0, 1)$ es simétrica respecto del origen (i.e., $z_{1-\gamma} = -z_\gamma$), tenemos que, para cada $\beta \in (0.5, 1)$,

1.

$$\mu_1(\mathbf{X}) = \bar{X} - \frac{\sigma}{\sqrt{n}}z_\beta$$

es una cota inferior de confianza de nivel β para μ ;

2.

$$\mu_2(\mathbf{X}) = \bar{X} + \frac{\sigma}{\sqrt{n}}z_\beta$$

es una cota superior de confianza de nivel β para μ ;

3.

$$I(\mathbf{X}) = \left[\bar{X} - \frac{\sigma}{\sqrt{n}}z_{(1+\beta)/2}, \bar{X} + \frac{\sigma}{\sqrt{n}}z_{(1+\beta)/2} \right]$$

es un intervalo de confianza de nivel β para μ . \square

2.3. Varianza de la normal con media conocida

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria $X \sim \mathcal{N}(\mu, \sigma^2)$, con media μ conocida. El estimador de máxima verosimilitud para σ^2 es

$$\widehat{\sigma^2}_{mv}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

Para construir un pivote para la varianza observamos que

$$\frac{n}{\sigma^2} \widehat{\sigma^2}_{mv}(\mathbf{X}) = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n Z_i^2,$$

donde $Z_i = \frac{X_i - \mu}{\sigma}$ son variables independientes cada una con distribución normal estándar $\mathcal{N}(0, 1)$. En otras palabras, la distribución de la variable aleatoria $\frac{n}{\sigma^2} \widehat{\sigma^2}_{mv}(\mathbf{X})$ coincide con la distribución de una suma de la forma $\sum_{i=1}^n Z_i^2$, donde las Z_i son $\mathcal{N}(0, 1)$ independientes. Por lo tanto,

$$Q(\mathbf{X}, \sigma^2) = \frac{n \widehat{\sigma^2}_{mv}(\mathbf{X})}{\sigma^2} \sim \chi_n^2$$

es un *pivote* para σ^2 .

Como el pivote para la varianza $Q(\mathbf{X}, \sigma^2)$ goza de las propiedades enunciadas en la sección 1.1.1 para pivotes decrecientes,

- la función de distribución de $Q(\mathbf{X}, \sigma^2)$ es continua y estrictamente creciente,
- para cada \mathbf{x} , la función $Q(\mathbf{x}, \sigma^2)$ es continua y monótona decreciente respecto de σ^2 ,

las cotas e intervalos de confianza para la varianza se pueden construir resolviendo la ecuación $Q(\mathbf{X}, \sigma^2) = \chi_{n, \gamma}^2$, donde $\chi_{n, \gamma}^2$ designa el cuantil- γ de la distribución chi cuadrado con n grados de libertad.

Observando que

$$Q(\mathbf{X}, \sigma^2) = \chi_{n, \gamma}^2 \iff \frac{n \widehat{\sigma^2}_{mv}(\mathbf{X})}{\sigma^2} = \chi_{n, \gamma}^2 \iff \sigma^2 = \frac{n \widehat{\sigma^2}_{mv}(\mathbf{X})}{\chi_{n-1, \gamma}^2},$$

se deduce que, para cada $\beta \in (0, 1)$,

1.

$$\sigma_1^2(\mathbf{X}) = \frac{n \widehat{\sigma^2}_{mv}(\mathbf{X})}{\chi_{n, \beta}^2}$$

es una cota inferior de confianza de nivel β para σ^2 ;

2.

$$\sigma_2^2(\mathbf{X}) = \frac{n \widehat{\sigma^2}_{mv}(\mathbf{X})}{\chi_{n, 1-\beta}^2}$$

es una cota superior de confianza de nivel β para σ^2 ;

3.

$$I(\mathbf{X}) = \left[\frac{n \widehat{\sigma^2}_{mv}(\mathbf{X})}{\chi_{n, (1+\beta)/2}^2}, \frac{n \widehat{\sigma^2}_{mv}(\mathbf{X})}{\chi_{n, (1-\beta)/2}^2} \right]$$

es un intervalo de confianza de nivel β para σ^2 . □

3. Intervalos aproximados para ensayos Bernoulli

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria $X \sim \text{Bernoulli}(p)$, donde $n \gg 1$. El estimador de máxima verosimilitud para p es

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Para construir un pivote para la varianza observamos que de acuerdo con el Teorema central del límite la distribución aproximada de $\sum_{i=1}^n X_i$ es una normal $\mathcal{N}(np, np(1-p))$ y en consecuencia

$$Q(\mathbf{X}, p) = \frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}} \sim \mathcal{N}(0, 1)$$

es un pivote asintótico para p .

Usando métodos analíticos se puede mostrar que $Q(\mathbf{X}, p)$ es una función continua y decreciente en $p \in (0, 1)$. Como el pivote asintótico para p goza de las propiedades enunciadas en la sección 1.1.1 para pivotes decrecientes, las cotas e intervalos de confianza para p se pueden construir resolviendo la ecuación $Q(\mathbf{X}, p) = z_\gamma$, donde z_γ designa el cuantil- γ de la distribución normal estándar $\mathcal{N}(0, 1)$.

Para resolver la ecuación $Q(\mathbf{X}, p) = z$ se elevan ambos miembros al cuadrado y se obtiene una ecuación cuadrática en p cuya solución es

$$p = \frac{z^2 + 2n\bar{X}}{2z^2 + 2n} \pm \frac{z\sqrt{z^2 + 4n\bar{X}(1 - \bar{X})}}{2z^2 + 2n}$$

Usando que la densidad de la distribución $\mathcal{N}(0, 1)$ es simétrica respecto del origen tenemos que, para cada $\beta \in (0.5, 1)$,

1.

$$p_1(\mathbf{X}) = \frac{z_\beta^2 + 2n\bar{X}}{2z_\beta^2 + 2n} - \frac{z_\beta\sqrt{z_\beta^2 + 4n\bar{X}(1 - \bar{X})}}{2z_\beta^2 + 2n}$$

es una cota inferior de confianza de nivel β para p ;

2.

$$p_2(\mathbf{X}) = \frac{z_\beta^2 + 2n\bar{X}}{2z_\beta^2 + 2n} + \frac{z_\beta\sqrt{z_\beta^2 + 4n\bar{X}(1 - \bar{X})}}{2z_\beta^2 + 2n}$$

es una cota superior de confianza de nivel β para p ;

3.

$$I(\mathbf{X}) = \left[\frac{z_{(1+\beta)/2}^2 + 2n\bar{X}}{2z_{(1+\beta)/2}^2 + 2n} \pm \frac{z_{(1+\beta)/2}\sqrt{z_{(1+\beta)/2}^2 + 4n\bar{X}(1 - \bar{X})}}{2z_{(1+\beta)/2}^2 + 2n} \right] \quad (10)$$

donde $[a \pm b] = [a - b, a + b]$, es un intervalo de confianza de nivel β para p . \square



Ejemplo 3.1 (Las agujas de Buffon). Se arroja al azar una aguja de longitud 1 sobre un plano dividido por rectas paralelas separadas por una distancia igual a 2.

Si localizamos la aguja mediante la distancia ρ de su centro a la recta más cercana y el ángulo agudo α entre la recta y la aguja, el espacio muestral es el rectángulo $0 \leq \rho \leq 1$ y $0 \leq \alpha \leq \pi/2$. El evento “la aguja interseca la recta” ocurre cuando $\rho \leq \frac{1}{2} \operatorname{sen} \alpha$ y su probabilidad es

$$p = \frac{\int_0^{\pi/2} \frac{1}{2} \operatorname{sen} \alpha d\alpha}{\pi/2} = \frac{1}{\pi}.$$

Con el objeto de estimar π se propone construir un intervalo de confianza de nivel $\beta = 0.95$ para p , basado en los resultados de realizar el experimentos de Buffon con $n = 100$ agujas.

Poniendo en (10) $n = 100$ y $z_{(1+\beta)/2} = z_{0.975} = 1.96$ se obtiene que

$$\begin{aligned} I(\mathbf{X}) &= \left[\frac{1.96^2 + 200\bar{X}}{2(1.96)^2 + 200} \pm \frac{1.96\sqrt{1.96^2 + 400\bar{X}(1 - \bar{X})}}{2(1.96)^2 + 200} \right] \\ &= \left[\frac{3.8416 + 200\bar{X}}{207.6832} \pm \frac{1.96\sqrt{3.8416 + 400\bar{X}(1 - \bar{X})}}{207.6832} \right] \end{aligned}$$

Al realizar el experimento se observó que 28 de las 100 agujas intersectaron alguna recta. Con ese dato el estimador de máxima verosimilitud para p es $\bar{X} = 0.28$ y en consecuencia se obtiene el siguiente intervalo de confianza para p

$$\begin{aligned} I(\mathbf{X}) &= \left[\frac{3.8416 + 200(0.28)}{207.6832} \pm \frac{1.96\sqrt{3.8416 + 400(0.28)(1 - 0.28)}}{207.6832} \right] \\ &= [0.28814 \pm 0.08674] = [0.20140, 0.37488]. \end{aligned}$$

De donde se obtiene la siguiente estimación: $2.66 \leq \pi \leq 4.96$. □

Nota Bene. Notando que la longitud del intervalo de confianza de nivel $\beta > 1/2$ para p se puede acotar de la siguiente forma

$$|I(\mathbf{X})| = \frac{z_{(1+\beta)/2}\sqrt{z_{(1+\beta)/2}^2 + 4n\bar{X}(1 - \bar{X})}}{z_{(1+\beta)/2}^2 + n} \leq \frac{z_{(1+\beta)/2}\sqrt{z_{(1+\beta)/2}^2 + n}}{z_{(1+\beta)/2}^2 + n} < \frac{z_{(1+\beta)/2}}{\sqrt{n}},$$

se puede mostrar que para garantizar que $|I(\mathbf{X})| < \epsilon$, donde ϵ es positivo y “pequeño” basta tomar $n \geq (z_{(1+\beta)/2}/\epsilon)^2$. □

Ejemplo 3.2 (Las agujas de Buffon (continuación)). ¿Cuántas agujas deben arrojarse si se desea estimar π utilizando un intervalo de confianza para p , de nivel 0.95, cuyo margen de error sea 0.01? De acuerdo con la observación anterior basta tomar $n \geq (1.96/0.01)^2 = 38416$.

Simulando 38416 veces el experimento de Buffon obtuvimos 12222 éxitos. Con ese dato el estimador de máxima verosimilitud para p es 0.31814... y el intervalo para p es

$$I(\mathbf{X}) = [0.31350, 0.32282].$$

De donde se obtiene la siguiente estimación: $3.09766 \leq \pi \leq 3.18969$. \square

4. Comparación de dos muestras normales

Supongamos que $\mathbf{X} = (X_1, \dots, X_m)$ es una muestra aleatoria de tamaño m de una distribución normal $\mathcal{N}(\mu_X, \sigma_X^2)$, y que $\mathbf{Y} = (Y_1, \dots, Y_n)$ es una muestra aleatoria de tamaño n de una distribución normal $\mathcal{N}(\mu_Y, \sigma_Y^2)$. Más aún, supongamos que las muestras \mathbf{X} e \mathbf{Y} son independientes. Usualmente los parámetros μ_X , μ_Y , σ_X^2 y σ_Y^2 son desconocidos.

4.1. Cotas e intervalos de confianza para la diferencia de medias

Queremos estimar $\Delta = \mu_X - \mu_Y$.

4.1.1. Varianzas conocidas

Para construir un pivote para la diferencia de medias, Δ , cuando las varianzas σ_X^2 y σ_Y^2 son conocidas, observamos que el estimador de máxima verosimilitud para $\Delta = \mu_X - \mu_Y$ es $\bar{X} - \bar{Y}$ y que

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\Delta, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right) \quad (11)$$

En consecuencia,

$$Q(\mathbf{X}, \mathbf{Y}, \Delta) = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0, 1), \quad (12)$$

es un pivote para la diferencia de medias Δ .

Como el pivote para la diferencia de medias, $Q(\mathbf{X}, \mathbf{Y}, \Delta)$, goza de las propiedades enunciadas en la sección 1.1.1 las cotas e intervalos de confianza para Δ se pueden construir resolviendo la ecuación $Q(\mathbf{X}, \mathbf{Y}, \Delta) = z_\gamma$, donde z_γ designa el cuantil- γ de la distribución $\mathcal{N}(0, 1)$. \square

4.1.2. Varianzas desconocidas.

Supongamos ahora que las varianzas σ_X^2 y σ_Y^2 son desconocidas. Hay dos posibilidades: las varianzas son iguales o las varianzas son distintas.

Caso 1: Varianzas iguales. Supongamos que $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. En tal caso

$$Z = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}} = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\sigma^2} \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \mathcal{N}(0, 1).$$

La varianza desconocida σ^2 se puede estimar ponderando “adecuadamente” los estimadores de varianza $S_X^2 = \frac{1}{m-1} \sum (X_i - \bar{X})^2$ y $S_Y^2 = \frac{1}{n-1} \sum (Y_j - \bar{Y})^2$,

$$S_P^2 := \frac{m-1}{m+n-2} S_X^2 + \frac{n-1}{m+n-2} S_Y^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}.$$

Se puede mostrar que

$$U := \frac{(n+m-2)}{\sigma^2} S_P^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \sim \chi_{n+m-2}.$$

Como las variables Z y U son independientes, se obtiene que

$$T = \frac{Z}{\sqrt{U/(m+n-2)}} = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{S_P^2} \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

Por lo tanto,

$$Q(\mathbf{X}, \mathbf{Y}, \Delta) = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{S_P^2} \sqrt{\frac{1}{m} + \frac{1}{n}}} \tag{13}$$

es un pivote para la diferencia de medias Δ . Debido a que el pivote goza de las propiedades enunciadas en la sección 1.1.1, las cotas e intervalos de confianza para Δ se pueden construir resolviendo la ecuación $Q(\mathbf{X}, \mathbf{Y}, \Delta) = t_{m+n-2, \gamma}$, donde $t_{m+n-2, \gamma}$ designa el cuantil- γ de la distribución t de Student con $m+n-2$ grados de libertad. \square

Caso 2: Varianzas distintas. En varios manuales de Estadística (el de Walpole, por ejemplo) se afirma que la distribución de la variable

$$Q(\mathbf{X}, \mathbf{Y}, \Delta) = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}$$

es una t de Student con ν grados de libertad, donde

$$\nu = \frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n} \right)^2}{\frac{\left(\frac{S_X^2}{m} \right)^2}{m-1} + \frac{\left(\frac{S_Y^2}{n} \right)^2}{n-1}}$$

Es de suponer que este “misterioso” valor de ν es el resultado de alguna controversia entre Estadísticos profesionales con suficiente experiencia para traducir semejante jeroglífico. Sin embargo, ninguno de los manuales se ocupa de revelar este misterio. \square

4.2. Cotas e intervalos de confianza para el cociente de varianzas.

Queremos estimar el cociente de las varianzas $R = \sigma_X^2/\sigma_Y^2$.

Si las medias μ_X y μ_Y son desconocidas, las varianzas σ_X^2 y σ_Y^2 se pueden estimar mediante sus estimadores insesgados $S_X^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2$ y $S_Y^2 = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y})^2$.

Debido a que las variables

$$U := \frac{(m-1)}{\sigma_X^2} S_X^2 \sim \chi_{m-1}^2 \quad \text{y} \quad V := \frac{(n-1)}{\sigma_Y^2} S_Y^2 \sim \chi_{n-1}^2$$

son independientes, tenemos que el cociente

$$\frac{U/(m-1)}{V/(n-1)} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{1}{R} \left(\frac{S_X^2}{S_Y^2} \right)$$

se distribuye como una F de Fisher con $m-1$ y $n-1$ grados de libertad.

Por lo tanto,

$$Q(\mathbf{X}, \mathbf{Y}, R) = \frac{1}{R} \left(\frac{S_X^2}{S_Y^2} \right) \sim F_{m-1, n-1}$$

es un pivote para el cociente de varianzas $R = \sigma_X^2/\sigma_Y^2$. Debido a que el pivote goza de las propiedades enunciadas en la sección 1.1.1, las cotas e intervalos de confianza para R se pueden construir resolviendo la ecuación $Q(\mathbf{X}, \mathbf{Y}, R) = F_{m-1, n-1, \gamma}$, donde $F_{m-1, n-1, \gamma}$ designa el cuantil- γ de la distribución F de Fisher con $m-1$ y $n-1$ grados de libertad. \square

5. Comparación de dos muestras

5.1. Planteo general

Supongamos que tenemos dos muestras aleatorias independientes $\mathbf{X} = (X_1, \dots, X_m)$ e $\mathbf{Y} = (Y_1, \dots, Y_n)$ con distribuciones dependientes de los parámetros ξ y η , respectivamente. Queremos estimar la diferencia

$$\Delta = \xi - \eta.$$

En lo que sigue mostraremos que, bajo ciertas hipótesis, podemos construir cotas e intervalos de confianza (aproximados) basados en el comportamiento de la diferencia $\hat{\xi}_m - \hat{\eta}_n$, donde $\hat{\xi}_m = \hat{\xi}(\mathbf{X})$ y $\hat{\eta}_n = \hat{\eta}(\mathbf{Y})$ son estimadores de los parámetros ξ y η , respectivamente.

En todo lo que sigue vamos a suponer que los estimadores $\hat{\xi}_m$ y $\hat{\eta}_n$ tienen la propiedad de normalidad asintótica. Esto es,

$$\begin{aligned} \sqrt{m}(\hat{\xi}_m - \xi) &\rightarrow \mathcal{N}(0, \sigma^2) && \text{cuando } m \rightarrow \infty, \\ \sqrt{n}(\hat{\eta}_n - \eta) &\rightarrow \mathcal{N}(0, \tau^2) && \text{cuando } n \rightarrow \infty, \end{aligned}$$

donde σ^2 y τ^2 pueden depender de ξ y η , respectivamente. Sea $N = m+n$ y supongamos que para algún $0 < \rho < 1$,

$$\frac{m}{N} \rightarrow \rho, \quad \frac{n}{M} \rightarrow 1-\rho \quad \text{cuando } m \text{ y } n \rightarrow \infty,$$

de modo que, cuando $N \rightarrow \infty$ tenemos

$$\sqrt{N}(\hat{\xi}_m - \xi) \rightarrow \mathcal{N}\left(0, \frac{\sigma^2}{\rho}\right) \quad \text{y} \quad \sqrt{N}(\hat{\eta}_n - \eta) \rightarrow \mathcal{N}\left(0, \frac{\tau^2}{1-\rho}\right).$$

Entonces, vale que

$$\sqrt{N} \left[(\hat{\xi}_m - \xi) - (\hat{\eta}_n - \eta) \right] \rightarrow \mathcal{N}\left(0, \frac{\sigma^2}{\rho} + \frac{\tau^2}{1-\rho}\right)$$

o, equivalentemente, que

$$\frac{(\hat{\xi}_m - \hat{\eta}_n) - \Delta}{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}} \rightarrow \mathcal{N}(0, 1) \tag{14}$$

Si σ^2 y τ^2 son conocidas, de (14) resulta que

$$Q(\mathbf{X}, \mathbf{Y}, \Delta) = \frac{(\hat{\xi}_m - \hat{\eta}_n) - \Delta}{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}} \tag{15}$$

es un pivote (aproximado) para la diferencia Δ .

Si σ^2 y τ^2 son desconocidas y $\widehat{\sigma^2}$ y $\widehat{\tau^2}$ son estimadores consistentes para σ^2 y τ^2 , se puede demostrar que la relación (14) conserva su validez cuando σ^2 y τ^2 se reemplazan por $\widehat{\sigma^2}$ y $\widehat{\tau^2}$, respectivamente y entonces

$$Q(\mathbf{X}, \mathbf{Y}, \Delta) = \frac{(\hat{\xi}_m - \hat{\eta}_n) - \Delta}{\sqrt{\frac{\widehat{\sigma^2}}{m} + \frac{\widehat{\tau^2}}{n}}} \tag{16}$$

es un pivote (aproximado) para la diferencia Δ .

Para mayores detalles se puede consultar el libro Lehmann, E. L. (1999) *Elements of Large-Sample Theory*. Springer, New York.

Nota Bene. Notar que el argumento anterior proporciona un método general de naturaleza asintótica. En otras palabras, en la práctica los resultados que se obtienen son aproximados. Dependiendo de los casos particulares existen diversos refinamientos que permiten mejorar esta primera aproximación.

5.2. Problema de dos muestras binomiales

Sean $\mathbf{X} = (X_1, \dots, X_m)$ e $\mathbf{Y} = (Y_1, \dots, Y_n)$ dos muestras aleatorias independientes de dos variables aleatorias X e Y con distribución Bernoulli de parámetros p_X y p_Y , respectivamente. Queremos estimar la diferencia

$$\Delta = p_X - p_Y$$

Para construir cotas e intervalos de confianza usaremos los estimadores de máxima verosimilitud para las probabilidades p_X y p_Y

$$\hat{p}_X = \bar{X} = \frac{1}{m} \sum_{i=1}^m X_i, \quad \hat{p}_Y = \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j,$$

Vamos a suponer que los volúmenes de las muestras, m y n , son suficientemente grandes y que ninguna de las dos variables está sobre representada (i.e. m y n son del mismo orden de magnitud).

Debido a que los estimadores \bar{X} y \bar{Y} son consistentes para las p_X y p_Y , resulta que los estimadores $\bar{X}(1 - \bar{X})$ y $\bar{Y}(1 - \bar{Y})$ son consistentes para las varianzas $p_X(1 - p_X)$ y $p_Y(1 - p_Y)$, respectivamente. Por lo tanto,

$$Q(\mathbf{X}, \mathbf{Y}, \Delta) = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{1}{m}\bar{X}(1 - \bar{X}) + \frac{1}{n}\bar{Y}(1 - \bar{Y})}} \quad (17)$$

es un pivote (aproximado) para Δ . \square

Ejemplo 5.1. Se toma una muestra aleatoria de 180 argentinos y resulta que 30 están desocupados. Se toma otra muestra aleatoria de 200 uruguayos y resulta que 25 están desocupados. ¿Hay evidencia suficiente para afirmar que la tasa de desocupación de la población Argentina es superior a la del Uruguay?

Solución. La población desocupada de la Argentina puede modelarse con una variable aleatoria $X \sim \text{Bernoulli}(p_X)$ y la del Uruguay con una variable aleatoria $Y \sim \text{Bernoulli}(p_Y)$.

Para resolver el problema utilizaremos una cota inferior de nivel de significación $\beta = 0.95$ para la diferencia $\Delta = p_X - p_Y$ basada en dos muestras aleatorias independientes \mathbf{X} e \mathbf{Y} de volúmenes $m = 180$ y $n = 200$, respectivamente.

En vista de que el pivote definido en (17) goza de las propiedades enunciadas en la sección 1.1.1, la cota inferior de nivel $\beta = 0.95$ para Δ se obtiene resolviendo la ecuación $Q(\mathbf{X}, \mathbf{Y}, \Delta) = z_{0.95}$.

Observando que

$$\begin{aligned} Q(\mathbf{X}, \mathbf{Y}, \Delta) = z_{0.95} &\iff \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{1}{180}\bar{X}(1 - \bar{X}) + \frac{1}{200}\bar{Y}(1 - \bar{Y})}} = 1.64 \\ &\iff \Delta = \bar{X} - \bar{Y} - 1.64\sqrt{\frac{1}{180}\bar{X}(1 - \bar{X}) + \frac{1}{200}\bar{Y}(1 - \bar{Y})} \end{aligned}$$

De acuerdo con los datos observados, $\bar{X} = \frac{30}{180} = \frac{1}{6}$ y $\bar{Y} = \frac{25}{200} = \frac{1}{8}$. Por lo tanto, la cota inferior para Δ adopta la forma

$$\Delta(\mathbf{x}, \mathbf{y}) = \frac{1}{6} - \frac{1}{8} - 1.64\sqrt{\frac{1}{180}\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) + \frac{1}{200}\left(\frac{1}{8}\right)\left(\frac{7}{8}\right)} = -0.0178\dots$$

De este modo se obtiene la siguiente estimación $p_X - p_Y > -0.0178$ y de allí no se puede concluir que $p_X > p_Y$. \square

6. Apéndice: Demostración del Teorema llave

6.1. Preliminares de Análisis y Álgebra

En la prueba del Teorema 2.1 se usarán algunas nociones de Álgebra Línea¹ y el Teorema de cambio de variables para la integral múltiple².

Teorema 6.1 (Cambio de variables en la integral múltiple). *Sea $f : \mathbb{R}^n \rightarrow \mathbb{R}$ una función integrable. Sea $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $g = (g_1, \dots, g_n)$ una aplicación biyectiva, cuyas componentes tienen derivadas parciales de primer orden continuas. Esto es, para todo $1 \leq i, j \leq n$, las funciones $\frac{\partial}{\partial y_j} g_i(\mathbf{y})$ son continuas. Si el Jacobiano de g es diferente de cero en casi todo punto, entonces,*

$$\int_A f(\mathbf{x}) d\mathbf{x} = \int_{g^{-1}(A)} f(g(\mathbf{y})) |J_g(\mathbf{y})| d\mathbf{y},$$

para todo conjunto abierto $A \subset \mathbb{R}^n$, donde $J_g(\mathbf{y}) = \det \left(\left(\frac{\partial g_i(\mathbf{y})}{\partial y_j} \right)_{i,j} \right)$.

El siguiente resultado, que caracteriza la distribución de un cambio de variables aleatorias, es una consecuencia inmediata del Teorema 6.1.

Corolario 6.2. *Sea \mathbf{X} un vector aleatorio n -dimensional con función densidad de probabilidad $f_{\underline{\mathbf{X}}}(\mathbf{x})$. Sea $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ una aplicación que satisface las hipótesis del Teorema 6.1. Entonces, el vector aleatorio $\mathbf{Y} = \varphi(\mathbf{X})$ tiene función densidad de probabilidad $f_{\mathbf{Y}}(\mathbf{y})$ de la forma:*

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\varphi^{-1}(\mathbf{y})) |J_{\varphi^{-1}}(\mathbf{y})|.$$

Demostración. Cualquiera sea el conjunto abierto A se tiene que

$$\mathbb{P}(\mathbf{Y} \in A) = \mathbb{P}(\varphi(\mathbf{X}) \in A) = \mathbb{P}(\mathbf{X} \in \varphi^{-1}(A)) = \int_{\varphi^{-1}(A)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Aplicando el Teorema 6.1 para $g = \varphi^{-1}$ se obtiene

$$\int_{\varphi^{-1}(A)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int_A f_{\mathbf{X}}(\varphi^{-1}(\mathbf{y})) |J_{\varphi^{-1}}(\mathbf{y})| d\mathbf{y}.$$

Por ende

$$\mathbb{P}(\mathbf{Y} \in A) = \int_A f_{\mathbf{X}}(\varphi^{-1}(\mathbf{y})) |J_{\varphi^{-1}}(\mathbf{y})| d\mathbf{y}.$$

Por lo tanto, el vector aleatorio \mathbf{Y} tiene función densidad de probabilidad de la forma $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\varphi^{-1}(\mathbf{y})) |J_{\varphi^{-1}}(\mathbf{y})|$. \square

¹La noción de base ortonormal respecto del producto interno canónico en \mathbb{R}^n y la noción de matriz ortogonal. Si lo desea, aunque no es del todo cierto, puede pensar que las matrices ortogonales corresponden a rotaciones espaciales.

²Sobre la nomenclatura: Los vectores de \mathbb{R}^n se piensan como vectores columna y se notarán en negrita $\mathbf{x} = [x_1 \dots x_n]^T$.

6.2. Lema previo

Observación 6.3. Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una distribución $\mathcal{N}(0, \sigma^2)$. Por independencia, la distribución conjunta de las variables X_1, \dots, X_n tiene función densidad de probabilidad de la forma

$$\begin{aligned} f(\mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}x_i^2\right) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right) \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}\|_2^2\right). \end{aligned}$$

De la observación anterior es claro que la distribución conjunta de las variables X_1, \dots, X_n es invariantes por rotaciones. Más concretamente vale el siguiente resultado:

Lema 6.4 (Isotropía). Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable $\mathcal{N}(0, \sigma^2)$ y sea $B \in \mathbb{R}^{n \times n}$ una matriz ortogonal, i.e. $B^T B = B B^T = I_n$. Si $\underline{\mathbf{X}} = [X_1 \dots X_n]^T$, entonces $\underline{\mathbf{Y}} = [Y_1 \dots Y_n]^T = B \underline{\mathbf{X}}$ tiene la misma distribución conjunta que $\underline{\mathbf{X}}$. En particular las variables aleatorias Y_1, \dots, Y_n son independientes y son todas $\mathcal{N}(0, \sigma^2)$.

Demostración. Es consecuencia inmediata del Teorema de cambio de variables para $\mathbf{y} = g(\mathbf{x}) = B\mathbf{x}$. Debido a que B es una matriz ortogonal, $g^{-1}(\mathbf{y}) = B^T \mathbf{y}$ y $J_{g^{-1}}(\mathbf{y}) = \det(B^T) = \pm 1$

$$\begin{aligned} f_{\underline{\mathbf{Y}}}(\mathbf{y}) &= f_{\underline{\mathbf{X}}}(B^T \mathbf{y}) |\det(B^T)| = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\|B^T \mathbf{y}\|_2^2\right) |\det(B^T)| \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}\|_2^2\right). \end{aligned}$$

En la última igualdad usamos que $\|B^T \mathbf{y}\|_2 = \|\mathbf{y}\|_2$ debido a que las transformaciones ortogonales preservan longitudes. \square

6.3. Demostración del Teorema.

Sin perder generalidad se puede suponer que $\mu = 0$. Sea $\mathcal{B} = \{b_1, b_2, \dots, b_n\}$ una base ortonormal de \mathbb{R}^n , donde $b_1 = \frac{1}{\sqrt{n}}[1 \dots 1]^T$. Sea $B \in \mathbb{R}^{n \times n}$ la matriz ortogonal cuya i -ésima fila es b_i^T . De acuerdo con el Lema 6.4 el vector aleatorio $\underline{\mathbf{Y}} = [Y_1 \dots Y_n]^T = B \underline{\mathbf{X}}$ tiene la misma distribución que $\underline{\mathbf{X}}$.

En primer lugar, observamos que

$$Y_1 = b_1^T \underline{\mathbf{X}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i = \sqrt{n}(\bar{X}).$$

En segundo lugar,

$$\sum_{i=1}^n Y_i^2 = \underline{\mathbf{Y}}^T \underline{\mathbf{Y}} = (B \underline{\mathbf{X}})^T B \underline{\mathbf{X}} = \underline{\mathbf{X}}^T B^T B \underline{\mathbf{X}} = \underline{\mathbf{X}}^T \underline{\mathbf{X}} = \sum_{i=1}^n X_i^2.$$

En consecuencia,

$$\sum_{i=2}^n Y_i^2 = \sum_{i=1}^n X_i^2 - Y_1^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n (X_i - \bar{X})^2.$$

Las variables Y_1, \dots, Y_n son independientes. Como $\sqrt{n}(\bar{X})$ depende de Y_1 , mientras que $\sum_{i=1}^n (X_i - \bar{X})^2$ depende de Y_2, \dots, Y_n , resulta que \bar{X} y S^2 son independientes (lo que prueba la parte (c)). Además, $\sqrt{n}(\bar{X}) = Y_1 \sim \mathcal{N}(0, \sigma^2)$, por lo tanto $Z = \frac{\sqrt{n}(\bar{X})}{\sigma} \sim \mathcal{N}(0, 1)$ (lo que prueba la parte (a)). La parte (b) se deduce de que

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=2}^n \left(\frac{Y_i}{\sigma} \right)^2 \sim \chi_{n-1}^2,$$

pues las $n-1$ variables $Y_2/\sigma, \dots, Y_n/\sigma$ son independientes y con distribución $\mathcal{N}(0, 1)$. \square

7. Bibliografía consultada

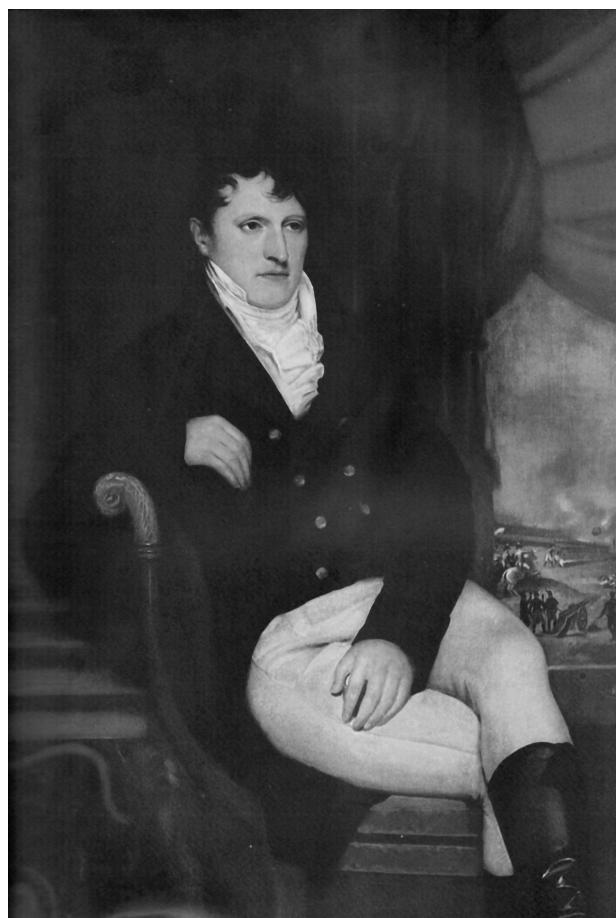
Para redactar estas notas se consultaron los siguientes libros:

1. Bolfarine, H., Sandoval, M. C.: Introdução à Inferência Estatística. SBM, Rio de Janeiro. (2001).
2. Borovkov, A. A.: Estadística matemática. Mir, Moscú. (1984).
3. Cramer, H.: Métodos matemáticos de estadística. Aguilar, Madrid. (1970).
4. Hoel P. G.: Introducción a la estadística matemática. Ariel, Barcelona. (1980).
5. Lehmann, E. L.: Elements of Large-Sample Theory. Springer, New York. (1999)
6. Maronna R.: Probabilidad y Estadística Elementales para Estudiantes de Ciencias. Editorial Exacta, La Plata. (1995).
7. Meyer, P. L.: Introductory Probability and Statistical Applications. Addison-Wesley, Massachusetts. (1972).
8. Walpole, R. E.: Probabilidad y estadística para ingenieros, 6a. ed., Prentice Hall, México. (1998)

Test de hipótesis y Test de bondad de ajuste (Borradores, Curso 23)

Sebastian Grynberg

3-12 de junio de 2013



*Que no se oiga ya que los ricos devoran a los pobres,
y que la justicia es sólo para los ricos.*
(Manuel Belgrano)

Índice

1. Planteo del problema	3
1.1. Test de hipótesis	3
1.2. Función de potencia	5
1.3. Nivel de significación	6
1.4. Sobre la construcción de reglas de decisión	7
2. Regiones de confianza y test de hipótesis	8
3. El método del pivote	9
3.1. Hipótesis fundamental simple contra alternativa bilateral	9
3.2. Hipótesis fundamental simple contra alternativa unilateral	10
3.3. Hipótesis fundamental unilateral contra alternativa unilateral	10
3.4. Algunos pivotes	11
4. Test para media de normales	13
4.1. Hipótesis sobre media con varianza conocida	13
4.2. Variaciones sobre el mismo tema	18
4.3. Hipótesis sobre media con varianza desconocida	20
5. Test para probabilidad de éxito de distribuciones Bernoulli	22
5.1. Test para moneda honesta (de lo simple a lo complejo)	23
5.2. Hipótesis fundamental simple	29
5.3. Hipótesis fundamental compuesta	32
6. Test para varianza de normales	34
6.1. Hipótesis sobre varianza con media conocida	34
6.2. Hipótesis sobre varianza con media desconocida	36
7. Comparación de dos muestras	37
7.1. Test para medias de dos muestras normales.	37
7.1.1. Varianzas conocidas	37
7.1.2. Varianzas desconocidas pero iguales.	37
7.2. Test F para varianzas de normales.	38
7.3. Planteo general	39
7.4. Problema de dos muestras binomiales	40
8. Test de la χ^2 para bondad de ajuste	42
8.1. Planteo del problema	42
8.2. Test de bondad de ajuste para hipótesis simples	43
8.3. Ejemplos (1 ^a parte)	45
8.4. Comentarios sobre el método	48
8.5. Test de bondad de ajuste para hipótesis compuestas	51

1. Planteo del problema

1.1. Test de hipótesis

Hipótesis estadística. El punto de partida es una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de una variable aleatoria X cuya función de distribución $F_X(x) = \mathbb{P}(X \leq x)$ pertenece a una familia paramétrica de distribuciones de probabilidad, $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.

En este contexto, una *hipótesis estadística* respecto de la distribución de probabilidades de la variable aleatoria X es una afirmación de la forma siguiente:

$$\text{“}F = F_\theta \text{ para algún } \theta \in \Theta_*\text{”}, \quad (1)$$

donde Θ_* es alguna parte del conjunto paramétrico Θ . Para simplificar la escritura, las hipótesis estadísticas (1) serán denotadas

$$H : \theta \in \Theta_*. \quad (2)$$

El problema general consiste en lo siguiente: en base a los resultados arrojados por la muestra aleatoria \mathbf{X} se quiere decidir entre dos hipótesis estadísticas sobre la distribución de probabilidades de la variable aleatoria X .

Test de hipótesis. Sean Θ_0 y Θ_1 dos subconjuntos del espacio paramétrico tales que $\Theta_0 \cap \Theta_1 = \emptyset$. El problema consiste en decidir entre las dos hipótesis

$$H_0 : \theta \in \Theta_0 \quad \text{contra} \quad H_1 : \theta \in \Theta_1,$$

basándose en el conocimiento de una muestra aleatoria, $\mathbf{X} = (X_1, \dots, X_n)$.

Como los valores de θ que no pertenecen a $\Theta_0 \cup \Theta_1$ no se examinan, se puede suponer que $\Theta = \Theta_0 \cup \Theta_1$, y que H_1 es la hipótesis *contraria* de H_0 . En tal caso, la hipótesis H_1 se puede escribir en la forma $H_1 : \theta \notin \Theta_0$. La hipótesis H_0 será llamada *hipótesis fundamental* o *hipótesis nula* y las hipótesis de la forma $H : \theta = \theta_1$, para $\theta_1 \in \Theta_1$, se llamarán *alternativas*.

Un *test* (o *regla de decisión*) para decidir entre las dos hipótesis H_0 contra H_1 es una aplicación medible $\delta : \mathbb{R}^n \rightarrow \{0, 1\}$ que le asigna a cada posible realización de la muestra aleatoria \mathbf{x} una y sólo una de las hipótesis. Concretamente, $\delta(\mathbf{X})$ es una variable aleatoria a valores en el $\{0, 1\}$. Cuando $\delta(\mathbf{X}) = 1$ se rechazará la hipótesis H_0 a favor de la hipótesis H_1 . En cambio, cuando $\delta(\mathbf{X}) = 0$ se aceptará la hipótesis H_0 .

Región crítica. Sea $\delta : \mathbb{R}^n \rightarrow \{0, 1\}$ un test para decidir entre las hipótesis H_0 contra H_1 . La región del espacio \mathbb{R}^n en la que $\delta(\mathbf{x}) = 1$:

$$\mathcal{R} := \{\mathbf{x} \in \mathbb{R}^n : \delta(\mathbf{x}) = 1\} \quad (3)$$

se denomina *región crítica* o *región de rechazo de la hipótesis fundamental*. La región crítica, \mathcal{R} , se identifica con la regla de decisión δ debido a que

$$\delta(\mathbf{x}) = \mathbf{1}\{\mathbf{x} \in \mathcal{R}\}. \quad (4)$$

Tipos de error. Todo test para decidir entre las hipótesis H_0 contra H_1 conduce a decisiones erróneas. Hay dos clases de decisiones erróneas.

- Las llamadas *errores de tipo I* que consisten en *RECHAZAR la hipótesis H_0 cuando ésta es verdadera*.
- Las llamadas *errores de tipo II* que consisten en *ACEPTAR la hipótesis H_0 cuando ésta es falsa*.

Nota Bene. Cuando $\theta \in \Theta_0$, la probabilidad de cometer un error de tipo I será

$$\mathbb{P}(\text{Rechazar } H_0 | \theta) = \mathbb{P}(\delta(\mathbf{X}) = 1 | \theta) = \mathbb{P}(\mathbf{X} \in \mathcal{R} | \theta).$$

Cuando $\theta \in \Theta_1$, la probabilidad de cometer un error de tipo II será

$$\mathbb{P}(\text{Aceptar } H_0 | \theta) = \mathbb{P}(\delta(\mathbf{X}) = 0 | \theta) = \mathbb{P}(\mathbf{X} \notin \mathcal{R} | \theta) = 1 - \mathbb{P}(\mathbf{X} \in \mathcal{R} | \theta).$$

Ejemplo 1.1. Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una distribución uniforme sobre el intervalo $(0, \theta)$, $\theta > 0$. Para decidir entre las dos hipótesis

$$H_0 : \theta \geq 2 \quad \text{contra} \quad H_1 : \theta < 2$$

consideramos el test $\delta(\mathbf{x}) = \mathbf{1}\{x_{(n)} \leq 3/2\}$, donde $x_{(n)} = \max(x_1, \dots, x_n)$ y queremos determinar, para cada $\theta > 0$, la probabilidad de decidir erróneamente.

Solución. Para calcular las probabilidades de decidir erróneamente estudiaremos la función $\beta : (0, \infty) \rightarrow [0, 1]$ definida por

$$\beta(\theta) = \mathbb{P}(\text{Rechazar } H_0 | \theta) = \mathbb{P}(\delta(\mathbf{X}) = 1 | \theta) = \mathbb{P}_\theta \left(X_{(n)} \leq \frac{3}{2} \right), \quad \theta > 0. \quad (5)$$

Sabemos que $Q(\mathbf{X}, \theta) = X_{(n)}/\theta$ es un pivote para θ y que su distribución tiene densidad de probabilidades $f_Q(q) = nq^{n-1} \mathbf{1}\{0 < q < 1\}$. En consecuencia,

$$\begin{aligned} \beta(\theta) &= \mathbb{P}_\theta \left(X_{(n)} \leq \frac{3}{2} \right) = \mathbb{P} \left(\frac{X_{(n)}}{\theta} \leq \frac{3}{2\theta} \right) = \int_0^{\min(1, \frac{3}{2\theta})} nq^{n-1} dq \\ &= \min \left(1, \frac{3}{2\theta} \right)^n = \mathbf{1} \left\{ 0 < \theta \leq \frac{3}{2} \right\} + \left(\frac{3}{2\theta} \right)^n \mathbf{1} \left\{ \theta > \frac{3}{2} \right\}. \end{aligned} \quad (6)$$

Por lo tanto,

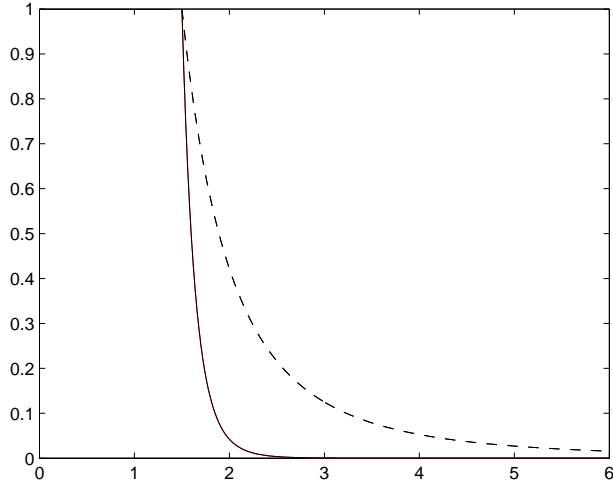


Figura 1: Gráfico de la función $\beta(\theta)$ para distintos volúmenes de muestra: en línea quebrada para volumen $n = 3$; en línea sólida para volumen $n = 11$. Notar que cuando n aumenta disminuyen las probabilidades de los errores de tipo I, pero aumentan las de los errores de tipo II.

- la probabilidad de que ocurra un error de tipo I cuando el verdadero valor del parámetro θ satisface $\theta \geq 2$ es $\beta(\theta) = \left(\frac{3}{2\theta}\right)^n$,
- la probabilidad de que ocurra un error de tipo II cuando el verdadero valor del parámetro θ satisface $\theta \in (0, 3/2]$ es $1 - \beta(\theta) = 1 - 1 = 0$,
- la probabilidad de que ocurra un error de tipo II cuando el verdadero valor del parámetro θ satisface $\theta \in (3/2, 2)$ es $1 - \beta(\theta) = 1 - \left(\frac{3}{2\theta}\right)^n$.

□

1.2. Función de potencia

La calidad de un test de hipótesis $\delta(\cdot)$ se caracteriza por el conjunto de probabilidades de decisiones erróneas (o riesgos de decisión).

Las probabilidades de los errores de un test $\delta(\cdot)$ se pueden representar en el gráfico de la función $\beta : \Theta \rightarrow [0, 1]$ definida por

$$\beta(\theta) := \mathbb{P}(\text{Rechazar } H_0 | \theta) = \mathbb{P}(\delta(\mathbf{X}) = 1 | \theta) = \mathbb{P}_\theta(\mathbf{X} \in \mathcal{R}), \quad (7)$$

llamada la *función de potencia* del test.¹

¹En control de calidad, a la función $\mathcal{L}(\theta) = 1 - \beta(\theta)$ se la llama *característica operativa* y su gráfico se llama la *curva característica operativa* del test.

En efecto, la probabilidad de que ocurra un error de tipo I cuando el verdadero valor del parámetro es $\theta \in \Theta_0$ será el valor de la probabilidad $\beta(\theta)$ y la probabilidad de cometer un error de tipo II cuando el verdadero valor del parámetro es $\theta \in \Theta_1$ será el valor de la probabilidad $1 - \beta(\theta)$.

Nota Bene. Una test puede considerarse “bueno” si los valores de su función de potencia están cerca del 0 en la región fundamental Θ_0 y cerca del 1 en la región alternativa Θ_1 . En general, establecido el volumen de la muestra, $\mathbf{X} = (X_1, \dots, X_n)$, no es posible construir test capaces de conciliar ambas exigencias.

1.3. Nivel de significación

Sea δ un test para decidir entre las hipótesis $H_0 : \theta \in \Theta_0$ contra $H_1 : \theta \in \Theta_1$. El número

$$\alpha(\delta) = \max_{\theta \in \Theta_0} \beta(\theta) \quad (8)$$

se llama *nivel de significación* del test. Dicho en palabras, el nivel de significación de un test es la máxima probabilidad de rechazar la hipótesis fundamental H_0 cuando ella es verdadera.

Ejemplo 1.2. Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una distribución $\mathcal{U}(0, \theta)$ y sea δ el test definido en el Ejemplo 1.1 para decidir entre las dos hipótesis $H_0 : \theta \geq 2$ contra $H_1 : \theta < 2$.

Debido a que la función de potencia $\beta(\theta)$ es decreciente en θ , el nivel de significación del test es

$$\alpha(\delta) = \max_{\theta \geq 2} \beta(\theta) = \beta(2) = \left(\frac{3}{4}\right)^n.$$

Para que, por ejemplo, el nivel de significación del test sea ≤ 0.05 , debe tomarse un volumen de muestra n tal que $(3/4)^n \leq 0.05$. Equivalentemente, $n \geq \log(0.05)/\log(3/4) = 10.413$. Para $n = 11$ el nivel del test resulta $\alpha(\delta) = 0.042\dots$ \square

Comentario sobre el nivel de significación. Utilizar un test de nivel de significación α significa que, en una larga serie de experimentos, no nos equivocaremos al rechazar la hipótesis H_0 , siendo que ella es verdadera, más que un $100\alpha\%$ de los casos. La elección del nivel de significación del test es arbitraria. Habitualmente, en calidad de α se elige alguno de los valores estándar, tales como 0.005, 0.01, 0.05, 0.1. Esta estandarización tiene la ventaja de que permite reducir el volumen de las tablas que se utilizan en el trabajo estadístico.

Nota Bene. La actitud que se tenga hacia la hipótesis fundamental antes de realizar el experimento es una circunstancia importante que puede influir en la elección del nivel de significación. Si se cree firmemente en su veracidad se necesitarán pruebas convincentes

en su contra para que se renuncie a ella. En tales condiciones hacen falta criterios de nivel α muy pequeños. Entonces, si la hipótesis fundamental es verdadera, la realización de un valor de muestra perteneciente a la región crítica \mathcal{R} será demasiado inverosímil. La concepción en la que se basa todo el razonamiento es la siguiente: si la probabilidad ϵ de cierto evento A es muy pequeña, consideramos prácticamente imposible el hecho de que este evento ocurra al realizar una sola prueba. Si ocurre, significa que su probabilidad no era tan pequeña.

Máxima potencia. Elegido el nivel de significación α del test de hipótesis, hay que prestarle atención a los valores de su función de *potencia* en la región alternativa Θ_1 . Si la potencia en Θ_1 resulta demasiado pequeña, los riesgos de cometer errores de tipo II son muy grandes y tal vez sea conveniente sustituir el nivel de significación por uno mayor. Entre todos los test de nivel α se prefieren aquellos que tengan la *potencia más alta* en toda la región alternativa Θ_1 .

1.4. Sobre la construcción de reglas de decisión

En la práctica, las reglas de decisión se construyen basándose en una estadística de la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$, i.e., son de la forma

$$\delta(\mathbf{X}) = \mathbf{1}\{T(\mathbf{X}) \in \mathcal{C}\}, \quad (9)$$

donde $T : \mathbb{R}^n \rightarrow \mathbb{R}$ es una función a valores reales y \mathcal{C} es una región de la recta real denominada la *región crítica* o *región de rechazo* del test: si $\delta(\mathbf{X}) = 1$ rechazamos la hipótesis H_0 y si $\delta(\mathbf{X}) = 0$ no la rechazamos.

Nota Bene. La estadística de la muestra, $T(\mathbf{X})$, con la que se construye la regla de decisión (9) debe contener toda la información relevante que hay en la muestra \mathbf{X} para reconstruir el parámetro θ sobre el que recaen las hipótesis H_0 y H_1 . Por ejemplo, si se hacen hipótesis sobre la media de la variable aleatoria X , es inútil observar simplemente todos los datos contenidos en la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$. Es intuitivamente claro que si se quiere tomar una decisión entre dos hipótesis sobre la media de una distribución hay que observar el promedio muestral $\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i$. Si la muestra es suficientemente grande, este valor se no puede desviar demasiado del verdadero valor de la media. Si el desvío fuese desconocido, para tener una idea de su tamaño bastará con observar el valor de la varianza muestral $S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Esos dos datos deberían ser suficientes para tomar una decisión sobre una hipótesis sobre la media.

Algunos problemas

1. Dado un test caracterizar su función de potencia, determinar su nivel y los distintos tipos de riesgos estadísticos.
2. Construcción de test prefijando el nivel α y el volumen de la muestra aleatoria n .

3. Construcción de test prefijando el nivel α y la potencia β en alguno de los parámetros alternativos.

Nota Bene. El objetivo de estas notas es presentar una introducción para tratar algunos problemas de carácter muy elemental y el modo de resolverlos mediante razonamientos intuitivos (lo más rigurosos posibles dentro del marco de un curso elemental).²

2. Regiones de confianza y test de hipótesis

Supongamos que disponemos de regiones de confianza $S(\mathbf{X})$ de nivel β para el parámetro θ y queremos construir un test para decidir entre las hipótesis

$$H_0 : \theta = \theta_0 \quad \text{contra} \quad H_1 : \theta \neq \theta_0.$$

Debido a que la región de confianza se construye con el objeto de capturar al verdadero valor del parámetro (con alta probabilidad de lograrlo) parece claro que si se observa un resultado \mathbf{x} tal que la región $S(\mathbf{x})$ contenga a θ_0 deberemos aceptar la hipótesis H_0 y rechazar la contraria H_1 . El argumento permite construir el siguiente test

$$\delta(\mathbf{X}) = \mathbf{1}\{S(\mathbf{X}) \ni \theta_0\}.$$

cuyo nivel de significación es

$$\alpha(\delta) = \mathbb{P}(\text{Rechazar } H_0 | \theta_0) = \mathbb{P}_{\theta_0}(S(\mathbf{X}) \ni \theta_0) = 1 - \mathbb{P}_{\theta_0}(S(\mathbf{X}) \not\ni \theta_0) = 1 - \beta.$$

□

Usando argumentos similares se obtienen los siguientes resultados.

1. Si $\theta_1(\mathbf{X})$ es una cota inferior de confianza de nivel $1 - \alpha$ para θ , entonces

$$\delta(\mathbf{X}) = \mathbf{1}\{\theta_0 < \theta_1(\mathbf{X})\}$$

es un test de nivel α para decidir entre las hipótesis

$$H_0 : \theta \leq \theta_0 \quad \text{contra} \quad H_1 : \theta > \theta_0.$$

2. Si $\theta_2(\mathbf{X})$ es una cota superior de confianza de nivel $1 - \alpha$ para θ , entonces

$$\delta(\mathbf{X}) = \mathbf{1}\{\theta_0 > \theta_2(\mathbf{X})\}$$

es un test de nivel α para decidir entre las hipótesis

$$H_0 : \theta \geq \theta_0 \quad \text{contra} \quad H_1 : \theta < \theta_0.$$

²Dependiendo de las normas de calidad que se le impongan al test y de la naturaleza de las hipótesis a ser confrontadas, existen metodologías generales para construir test óptimos que pueden consultarse en cualquier libro de *Estadística matemática*. Una exposición rigurosa puede encontrarse en el libro de Borovkov.

3. Si $[\theta_1(\mathbf{X}), \theta_2(\mathbf{X})]$ es un intervalo de confianza de nivel $1 - \alpha$ para θ . Entonces

$$\delta(\mathbf{X}) = \mathbf{1}\{[\theta_1(\mathbf{X}), \theta_2(\mathbf{X})] \not\ni \theta_0\}$$

es un test de nivel α para decidir entre las hipótesis

$$H_0 : \theta = \theta_0 \quad \text{contra} \quad H_1 : \theta \neq \theta_0.$$

Nota Bene. Notar que en cualquiera de los tres casos se rechaza la hipótesis H_0 cuando y solo cuando los intervalos de confianza están contenidos en la hipótesis alternativa H_1 .

3. El método del pivote

Cuando se quieren construir test de hipótesis para el parámetro desconocido θ lo más natural es comenzar la construcción apoyándose en algún estimador puntual del parámetro $\hat{\theta}(\mathbf{X})$ (cuya distribución depende de θ). El *método del pivote* consiste en transformar el estimador $\hat{\theta}(\mathbf{X})$ en un pivote $Q(\hat{\theta}(\mathbf{X}), \theta)$ y utilizarlo para construir el test deseado.

Nota Bene. Por definición, la distribución del pivote $Q(\hat{\theta}(\mathbf{X}), \theta)$ no depende de θ . Para cada $\gamma \in (0, 1)$ notaremos mediante q_γ el cuantil- γ del pivote.

En todo lo que sigue vamos a suponer que $Q(\hat{\theta}(\mathbf{X}), \theta)$ es un pivote que goza de las siguientes propiedades:

1. La función de distribución de $Q(\hat{\theta}(\mathbf{X}), \theta)$ es continua y estrictamente creciente.
2. La función $Q(t, \theta)$ es monótona decreciente en θ :

$$\theta_1 < \theta_2 \implies Q(t, \theta_1) > Q(t, \theta_2). \quad (10)$$

3.1. Hipótesis fundamental simple contra alternativa bilateral

Se desea un test de nivel α para decidir entre las hipótesis

$$H_0 : \theta = \theta_0 \quad \text{contra} \quad H_1 : \theta \neq \theta_0.$$

Proponemos un test de la forma

$$\delta(\mathbf{X}) = \mathbf{1}\left\{Q(\hat{\theta}(\mathbf{X}), \theta_0) < q_{\gamma_1}\right\} + \mathbf{1}\left\{Q(\hat{\theta}(\mathbf{X}), \theta_0) > q_{\gamma_2}\right\} \quad (11)$$

Como la hipótesis fundamental es de la forma $\theta = \theta_0$ el nivel de significación del test es

$$\begin{aligned} \alpha(\delta) &= \beta(\theta_0) = \mathbb{P}(\text{Rechazar } H_0 | \theta_0) = \mathbb{P}(Q(\hat{\theta}(\mathbf{X}), \theta_0) < q_{\gamma_1}) + \mathbb{P}(Q(\hat{\theta}(\mathbf{X}), \theta_0) > q_{\gamma_2}) \\ &= \mathbb{P}(Q(\hat{\theta}(\mathbf{X}), \theta_0) \leq q_{\gamma_1}) + 1 - \mathbb{P}(Q(\hat{\theta}(\mathbf{X}), \theta_0) \leq q_{\gamma_2}) = \gamma_1 + 1 - \gamma_2. \end{aligned}$$

Poniendo $\gamma_1 = \alpha/2$ y $\gamma_2 = 1 - \alpha/2$ obtenemos que $\alpha(\delta) = \alpha$. Por lo tanto, el test de hipótesis deseado puede obtenerse de la siguiente manera:

$$\delta(\mathbf{X}) = \mathbf{1}\left\{Q(\hat{\theta}(\mathbf{X}), \theta_0) < q_{\alpha/2}\right\} + \mathbf{1}\left\{Q(\hat{\theta}(\mathbf{X}), \theta_0) > q_{1-\alpha/2}\right\}. \quad (12)$$

□

3.2. Hipótesis fundamental simple contra alternativa unilateral

Se desea un test de nivel α para decidir entre las hipótesis

$$H_0 : \theta = \theta_0 \quad \text{contra} \quad H_1 : \theta > \theta_0.$$

Proponemos un test de la forma

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ Q(\hat{\theta}(\mathbf{X}), \theta_0) > q_\gamma \right\} \quad (13)$$

Como la hipótesis fundamental es de la forma $\theta = \theta_0$ el nivel de significación del test es

$$\alpha(\delta) = \beta(\theta_0) = \mathbb{P}(\text{Rechazar } H_0 | \theta_0) = \mathbb{P}\left(Q(\hat{\theta}(\mathbf{X}), \theta_0) > q_\gamma\right) = 1 - \gamma.$$

Poniendo $\gamma = 1 - \alpha$ obtenemos que $\alpha(\delta) = \alpha$. Por lo tanto, el test deseado puede obtenerse de la siguiente manera:

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ Q(\hat{\theta}(\mathbf{X}), \theta_0) > q_{1-\alpha} \right\}. \quad (14)$$

□

3.3. Hipótesis fundamental unilateral contra alternativa unilateral

1.- Como consecuencia de que la función $Q(t, \theta)$ es decreciente en θ , el test definido en (14) también se puede utilizar como test de nivel α para decidir entre las hipótesis

$$H_0 : \theta \leq \theta_0 \quad \text{contra} \quad H_1 : \theta > \theta_0.$$

En efecto, si $\theta \leq \theta_0$, entonces $Q(\hat{\theta}(\mathbf{X}), \theta) \geq Q(\hat{\theta}(\mathbf{X}), \theta_0)$ y en consecuencia

$$\beta(\theta) = \mathbb{P}(\text{Rechazar } H_0 | \theta) = \mathbb{P}_\theta\left(Q(\hat{\theta}(\mathbf{X}), \theta_0) > q_{1-\alpha}\right) \leq \mathbb{P}_\theta\left(Q(\hat{\theta}(\mathbf{X}), \theta) > q_{1-\alpha}\right) = \alpha.$$

Por lo tanto,

$$\max_{\theta \leq \theta_0} \beta(\theta) \leq \alpha.$$

Pero como $\beta(\theta_0) = \mathbb{P}_{\theta_0}\left(Q(\hat{\theta}(\mathbf{X}), \theta_0) > q_{1-\alpha}\right) = \alpha$, resulta que

$$\max_{\theta \leq \theta_0} \beta(\theta) = \alpha.$$

□

2.- Si se desea un test de nivel α para decidir entre las hipótesis

$$H_0 : \theta \geq \theta_0 \quad \text{contra} \quad H_1 : \theta < \theta_0$$

basta considerar

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ Q(\hat{\theta}(\mathbf{X}), \theta_0) < q_\alpha \right\}. \quad (15)$$

En efecto, si $\theta \geq \theta_0$, entonces $Q(\hat{\theta}(\mathbf{X}), \theta) \leq Q(\hat{\theta}(\mathbf{X}), \theta_0)$ y en consecuencia

$$\beta(\theta) = \mathbb{P}(\text{Rechazar } H_0 | \theta) = \mathbb{P}_\theta \left(Q(\hat{\theta}(\mathbf{X}), \theta_0) < q_\alpha \right) \leq \mathbb{P}_\theta \left(Q(\hat{\theta}(\mathbf{X}), \theta) < q_\alpha \right) = \alpha.$$

Por lo tanto,

$$\max_{\theta \geq \theta_0} \beta(\theta) \leq \alpha.$$

Pero como $\beta(\theta_0) = \mathbb{P}_{\theta_0}(Q(\hat{\theta}(\mathbf{X}), \theta_0) < q_\alpha) = \alpha$, resulta que

$$\max_{\theta \geq \theta_0} \beta(\theta) = \alpha.$$

□

3.4. Algunos pivotes

1. **Para media de normales con varianza conocida.** Si X_1, \dots, X_n es una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$, con σ^2 conocida, entonces

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

es un pivote para μ .

2. **Para media de normales con varianza desconocida.** Si X_1, \dots, X_n es una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$, con σ^2 desconocida, entonces

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

es un pivote para μ .

3. **Para varianza de normales con media conocida.** Si X_1, \dots, X_n es una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$, con μ conocida, entonces

$$\frac{n}{\sigma^2} \widehat{\sigma}_{mv}^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \chi_n^2$$

es un pivote para σ^2 .

4. **Para varianza de normales con media desconocida.** Si X_1, \dots, X_n es una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$, con μ desconocida, entonces

$$\frac{(n-1)}{\sigma^2} S^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2$$

es un pivote para σ^2 .

5. **Para probabilidad de éxito de distribuciones Bernoulli.** Si X_1, \dots, X_n es una m.a. de una distribución Bernoulli(p) y $n \gg 1$, entonces

$$\frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}} \sim \mathcal{N}(0, 1)$$

es un pivote aproximado para p .

6. **Para intensidad de exponenciales.** Si X_1, \dots, X_n es una m.a. de una distribución Exponencial(λ), entonces

$$2\lambda n \bar{X} = \lambda \sum_{i=1}^n X_i \sim \chi_{2n}^2$$

es un pivote para λ .

7. **Para extremo derecho de uniformes.** Si X_1, \dots, X_n es una m.a. de una distribución $\mathcal{U}(0, \theta)$, entonces

$$\frac{X_{(n)}}{\theta} = \frac{\max(X_1, \dots, X_n)}{\theta}$$

es un pivote para θ cuya densidad es $f(x) = nx^{n-1} \mathbf{1}\{0 \leq x \leq 1\}$.

8. **Para diferencia de medias de normales con varianzas conocidas.** Si X_1, \dots, X_m e Y_1, \dots, Y_n son dos m.a. independientes de distribuciones $\mathcal{N}(\mu_X, \sigma_X^2)$ y $\mathcal{N}(\mu_Y, \sigma_Y^2)$, con σ_X^2 y σ_Y^2 conocidas, entonces

$$\frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0, 1)$$

es un pivote para la diferencia de medias $\Delta = \mu_X - \mu_Y$.

9. **Para diferencia de medias de normales con varianzas desconocidas pero iguales.** Si X_1, \dots, X_m e Y_1, \dots, Y_n son dos m.a. independientes de distribuciones $\mathcal{N}(\mu_X, \sigma^2)$ y $\mathcal{N}(\mu_Y, \sigma^2)$, con varianza común σ^2 desconocida, entonces³

$$\frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{S_P^2} \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

³

$$S_P^2 := \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

es un pivote para la diferencia de medias $\Delta = \mu_X - \mu_Y$.

10. **Para cociente de varianzas de normales con medias desconocidas.** Si X_1, \dots, X_m e Y_1, \dots, Y_n son dos m.a. independientes de distribuciones $\mathcal{N}(\mu_X, \sigma_X^2)$ y $\mathcal{N}(\mu_Y, \sigma_Y^2)$, con μ_X y μ_Y desconocidas, entonces

$$\frac{1}{R} \left(\frac{S_X^2}{S_Y^2} \right) \sim F_{m-1, n-1}$$

es un pivote para el cociente de las varianzas $R = \sigma_X^2 / \sigma_Y^2$.

11. **Para diferencia de probabilidades de éxito de Bernoulli.** Si X_1, \dots, X_m e Y_1, \dots, Y_n son dos m.a. independientes de distribuciones Bernoulli(p_X) y Bernoulli(p_Y). Entonces,

$$\frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{1}{m}\bar{X}(1-\bar{X}) + \frac{1}{n}\bar{Y}(1-\bar{Y})}} \sim \mathcal{N}(0, 1)$$

es un pivote aproximado para la diferencia $\Delta = p_X - p_Y$.

4. Test para media de normales

En esta sección usaremos el método del pivote para construir test de hipótesis sobre la media de distribuciones normales.

4.1. Hipótesis sobre media con varianza conocida

Basados en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de una distribución normal $\mathcal{N}(\mu, \sigma^2)$ con varianza σ^2 conocida queremos construir un test de nivel de significación α para decidir entre las hipótesis

$$H_0 : \mu = \mu_0 \quad \text{contra} \quad H_1 : \mu \neq \mu_0,$$

donde μ_0 es un algún valor determinado.

Test de hipótesis

Para distribuciones normales con varianza conocida sabemos que

$$Q(\bar{X}, \mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

es un pivote para μ basado en $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Es fácil ver que el pivote satisface las dos condiciones enunciadas al principio de la Sección 3. De acuerdo con los resultados expuestos en la sección 3.1

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} < z_{\alpha/2} \right\} + \mathbf{1} \left\{ \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > z_{1-\alpha/2} \right\}, \quad (16)$$

es un test de nivel α para decidir entre las hipótesis $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$. Dicho en palabras, el test consiste en rechazar H_0 si $\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} < z_{\alpha/2}$ o $\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > z_{1-\alpha/2}$ y aceptarla en otro caso.

Nota Bene. Construir un test es la primera fase para decidir entre dos hipótesis. Construido el test es “obligatorio” analizar los riesgos de tomar decisiones erróneas. En otras palabras, el test debe acompañarse con su correspondiente función de potencia.

Función de potencia

Los riesgos de tomar decisiones erróneas utilizando el test de hipótesis definido en (16) pueden evaluarse caracterizando su correspondiente función de potencia: $\beta(\mu) := \mathbb{P}(\text{Rechazar } H_0 | \mu)$. Se trata de obtener una expresión “analítica” que nos permita caracterizar cuantitativa y cualitativamente las propiedades de dicha función.

Vale que

$$\beta(\mu) = \Phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) + \Phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}\right). \quad (17)$$

En efecto,

$$\begin{aligned} \beta(\mu) &= \mathbb{P}(\text{Rechazar } H_0 | \mu) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} < z_{\alpha/2}\right) + \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > z_{1-\alpha/2}\right) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} < z_{\alpha/2}\right) \\ &\quad + \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} > z_{1-\alpha/2}\right) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} < z_{\alpha/2} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) \\ &\quad + \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > -z_{\alpha/2} - \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) \\ &= \Phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) + \Phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}\right). \end{aligned}$$

Notar que la función de potencia dada en (17) satisface las siguientes propiedades

- (a) $\beta(\mu)$ es simétrica con respecto a μ_0 : $\beta(\mu_0 + m) = \beta(\mu_0 - m)$ para todo $m > 0$.
- (b) $\beta(\mu)$ es creciente⁴ sobre la semi-recta (μ_0, ∞) .
- (c) $\beta(\mu_0) = \alpha$.

⁴Derivar con respecto de μ la expresión (17) y hacer cuentas.

$$(d) \lim_{\mu \uparrow +\infty} \beta(\mu) = 1$$

Esto significa que a medida que nos alejamos de la hipótesis $\mu = \mu_0$ disminuye el riesgo de aceptar dicha hipótesis cuando es falsa. La forma típica del gráfico de la función de potencia correspondiente al test de la forma (16) para decidir entre las hipótesis $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$ puede observarse en las Figuras 2 y 3.

Nota Bene. La función de potencia es útil para determinar cuan grande debe ser la muestra aleatoria para conseguir ciertas especificaciones relativas a los errores de tipo II. Por ejemplo, supongamos que queremos determinar el volumen de la muestra n necesario para asegurar que la probabilidad de rechazar $H_0 : \mu = \mu_0$ cuando el verdadero valor de la media es μ_1 sea aproximadamente β . Esto es, queremos determinar n tal que

$$\beta(\mu_1) \approx \beta.$$

De la expresión (17), esto es equivalente a

$$\Phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) + \Phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}\right) \approx \beta. \quad (18)$$

Aunque la ecuación (18) no se pueda resolver analíticamente, se puede conseguir una solución aproximada mediante la siguiente observación.

1. Supongamos que $\mu_1 > \mu_0$. En tal caso, el primer término del lado izquierdo de (18) es despreciable, (es fácil ver que está acotado por $\alpha/2 \approx 0$) y por lo tanto, el problema se reduce a resolver la ecuación aproximada

$$\Phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}\right) \approx \beta.$$

En consecuencia, basta tomar n tal que $z_{\alpha/2} + \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} \approx z_\beta$ ó lo que es equivalente

$$n \approx \left(\frac{\sigma(z_\beta - z_{\alpha/2})}{\mu_1 - \mu_0} \right)^2. \quad (19)$$

2. Supongamos que $\mu_1 < \mu_0$. En tal caso, el segundo término del lado izquierdo de (18) es despreciable, y por lo tanto, el problema se reduce a resolver la ecuación aproximada

$$\Phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma}\right) \approx \beta.$$

En consecuencia, basta tomar n tal que

$$n \approx \left(\frac{\sigma(z_\beta - z_{\alpha/2})}{\mu_0 - \mu_1} \right)^2. \quad (20)$$

El resultado obtenido en (19) coincide con el resultado obtenido en (20) y es una aproximación razonable para el volumen de muestra necesario para asegurar que el error de tipo II en el valor $\mu = \mu_1$ es aproximadamente igual a $1 - \beta$.

Ejemplo 4.1. Si se envía una señal de valor μ desde un sitio A , el valor recibido en el sitio B se distribuye como una normal de media μ y desvío estándar 2. Esto es, el ruido que perturba la señal es una variable aleatoria $\mathcal{N}(0, 4)$. El receptor de la señal en el sitio B tiene suficientes motivos para sospechar que recibirá una señal de valor $\mu = 8$. Analizar la consistencia de dicha hipótesis suponiendo que la misma señal fue enviada en forma independientemente 5 veces desde el sitio A y el promedio del valor recibido en el sitio B es $\bar{X} = 9.5$.

Solución. Se trata de construir un test de hipótesis para decidir entre las hipótesis

$$H_0 : \mu = 8 \quad \text{contra} \quad H_1 : \mu \neq 8,$$

usando una muestra $\mathbf{X} = (X_1, \dots, X_5)$ de una distribución $\mathcal{N}(\mu, 4)$.

Test de hipótesis. Para un nivel de significación del 5% el test es de la forma

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ \left| \frac{\sqrt{5}(\bar{X} - 8)}{2} \right| > 1.96 \right\} \quad (21)$$

Decisión basada en la muestra observada. Calculamos el valor

$$\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \right| = \left| \frac{\sqrt{5}(9.5 - 8)}{2} \right| = 1.68$$

Como este valor es menor que $z_{1-\alpha/2} = z_{0.975} = 1.96$, se acepta la hipótesis $\mu = 8$. En otras palabras, los datos no son inconsistentes con la hipótesis $\mu = 8$.

Nota Bene. Notar que, si se relaja el nivel de significación al 10%, entonces la hipótesis $\mu = 8$ debe rechazarse debido a que el valor $z_{0.95} = 1.645$ es menor que 1.68.

Función de potencia. La función de potencia es

$$\beta(\mu) = \Phi \left(-1.96 + \frac{\sqrt{5}(8 - \mu)}{2} \right) + \Phi \left(-1.96 + \frac{\sqrt{5}(\mu - 8)}{2} \right). \quad (22)$$

Si se quiere determinar la probabilidad de cometer un error de tipo II cuando el valor real enviado es 10 basta poner $\mu = 10$ en la expresión (22) y calcular $1 - \beta(10)$:

$$1 - \Phi \left(-1.96 - \sqrt{5} \right) - \Phi \left(-1.96 + \sqrt{5} \right) = \Phi(-0.276) - \Phi(-4.196) = 0.392.$$

□

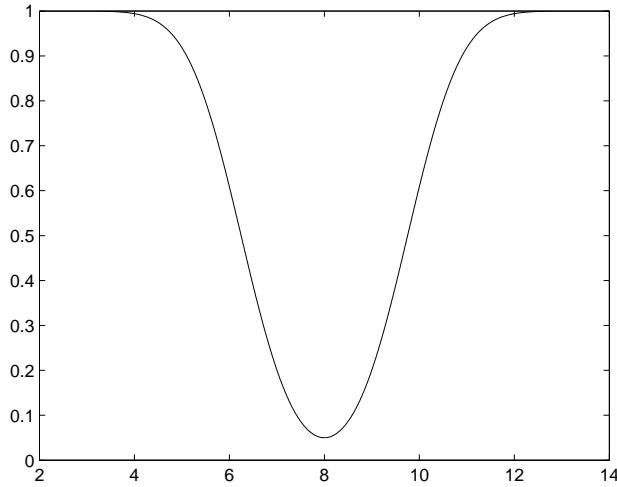


Figura 2: Gráfico de la función de potencia (22) correspondiente al test de hipótesis definido en (21) para decidir entre $H_0 : \mu = 8$ contra $H_1 : \mu \neq 8$ con un nivel de significación del 5% y basado en una muestra de volumen 5.

Ejemplo 4.2. Volvamos al problema del **Ejemplo 4.1**. Cuántas señales deberían enviarse para que el test de nivel de significación $\alpha = 0.05$ para $H_0 : \mu = 8$ contra $H_1 : \mu \neq 8$ tenga al menos una probabilidad igual a 0.75 de rechazar esa hipótesis cuando $\mu = 9.2$?

Solución. Como $z_{0.025} = -1.96$ y $z_{0.75} = 0.67$, de (19) resulta $n \approx \left(\frac{2(0.67+1.96)}{9.2-8} \right)^2 = 19.21$.

Para una muestra de volumen 20 el test adopta la forma

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ \left| \frac{\sqrt{20}(\bar{X} - 8)}{2} \right| > 1.96 \right\} = \mathbf{1} \left\{ \left| \sqrt{5}(\bar{X} - 8) \right| > 1.96 \right\} \quad (23)$$

y su función de potencia adopta la expresión

$$\beta(\mu) = \Phi(-1.96 + \sqrt{5}(8 - \mu)) + \Phi(-1.96 + \sqrt{5}(\mu - 8)). \quad (24)$$

En consecuencia,

$$\beta(9.2) = \Phi(-4.6433) + \Phi(0.72328) = 0.76525.$$

Dicho en palabras, si el mensaje se envía 20 veces, entonces hay un 76.52 % de posibilidades de que la hipótesis nula $\mu = 8$ sea rechazada cuando la media verdadera es 9.2.

□

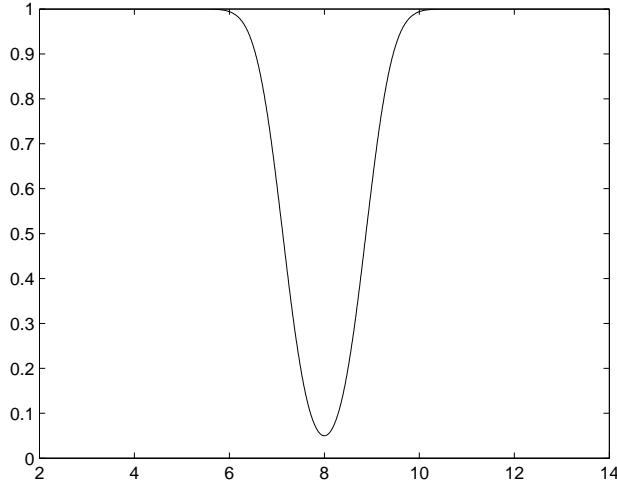


Figura 3: Gráfico de la función de potencia (24) correspondiente al test definido en (23) para decidir entre las hipótesis $H_0 : \mu = 8$ contra $H_1 : \mu \neq 8$ con un nivel de significación del 5 % y basado en una muestra de volumen 20.

Nota Bene. Comparando las Figuras 2 y 3 se puede ver que, fijado el nivel de significación del test, cuando se aumenta el volumen de la muestra disminuyen los errores de tipo II.

4.2. Variaciones sobre el mismo tema

Basados en una muestra $\mathbf{X} = (X_1, \dots, X_n)$ de una distribución normal $\mathcal{N}(\mu, \sigma^2)$ con varianza σ^2 conocida se quiere construir un test de nivel de significación α para decidir entre las hipótesis

$$H_0 : \mu = \mu_0 \quad \text{contra} \quad H_1 : \mu > \mu_0,$$

donde μ_0 es un algún valor determinado.

Usando los resultados expuestos en la sección 3.2 tenemos que

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > z_{1-\alpha} \right\}. \quad (25)$$

es un test de nivel α para decidir entre $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$. Dicho en palabras, el test de hipótesis consiste en rechazar H_0 si $\bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}}z_{1-\alpha}$ y aceptarla en otro caso.

Función de potencia. La función de potencia correspondiente al test (25) es

$$\begin{aligned}
\beta(\mu) &= \mathbb{P}(\text{Rechazar } H_0 | \mu) = \mathbb{P}_\mu \left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > z_{1-\alpha} \right) \\
&= \mathbb{P}_\mu \left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} > z_{1-\alpha} \right) \\
&= \mathbb{P}_\mu \left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > -z_\alpha - \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right) \\
&= \Phi \left(z_\alpha + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right). \tag{26}
\end{aligned}$$

De las propiedades de la función $\Phi(\cdot)$ y de la expresión (26) para la función de potencia se deduce que

- (a) $\beta(\mu)$ creciente.
- (b) $\beta(\mu_0) = \alpha$
- (c) $\lim_{\mu \uparrow +\infty} \beta(\mu) = 1$ y $\lim_{\mu \downarrow -\infty} \beta(\mu) = 0$.

Debido a que la función de potencia (26) es creciente, el test definido en (25) también se puede usar para decidir, con un nivel de significación α , entre la hipótesis

$$H_0 : \mu \leq \mu_0 \quad \text{contra} \quad H_1 : \mu > \mu_0.$$

Ejemplo 4.3. Volvamos al problema presentado en el **Ejemplo 4.1** pero supongamos que esta vez estamos interesados en testear con nivel de significación, $\alpha = 0.05$, la hipótesis $H_0 : \mu \leq 8$ contra la hipótesis alternativa $H_1 : \mu > 8$. (Recordar que disponemos de muestra aleatoria de volumen 5 de una población normal $\mathcal{N}(\mu, 4)$ cuyo promedio resultó ser $\bar{X} = 9.5$)

En este caso, el test de hipótesis definido en (25) puede enunciarse de la siguiente manera:

Rechazar H_0 cuando $\bar{X} > 8 + \frac{2}{\sqrt{5}}z_{0.95} = 9.4712$ y aceptarla en otro caso. (27)

Si se observó que $\bar{X} = 9.5$, entonces debe rechazarse la hipótesis $\mu \leq 8$ a favor de la alternativa $\mu > 9$. La función de potencia correspondiente al test de hipótesis (27) es

$$\beta(\mu) = \Phi \left(-1.64 + \frac{\sqrt{5}(\mu - 8)}{2} \right) \tag{28}$$

Si se quiere determinar la probabilidad de aceptar la hipótesis $\mu \leq 8$ cuando el valor real enviado es $\mu = 10$ basta poner $\mu = 10$ en la expresión (28) y calculamos:

$$1 - \beta(10) = 1 - \Phi \left(-1.64 + \sqrt{5} \right) = 0.27... \tag{29}$$

□

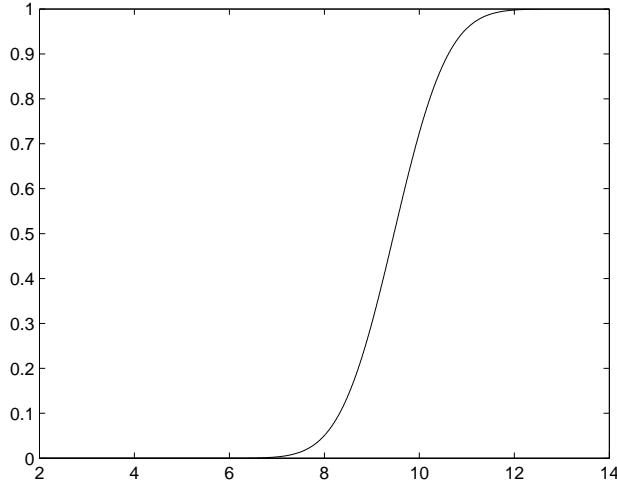


Figura 4: Gráfico de la función de potencia (28) correspondiente al test definido en (27) para decidir entre las hipótesis $H_0 : \mu \leq 8$ contra $H_1 : \mu > 8$ con un nivel de significación del 5 % y basado en una muestra de volumen 5.

4.3. Hipótesis sobre media con varianza desconocida

Basados en una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de una distribución normal $\mathcal{N}(\mu, \sigma^2)$ queremos construir un test de nivel de significación α para decidir entre las hipótesis

$$H_0 : \mu = \mu_0 \quad \text{contra} \quad H_1 : \mu \neq \mu_0,$$

donde μ_0 es un algún valor determinado.

Test de hipótesis

Para distribuciones normales sabemos que

$$Q(\bar{X}, \mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

es un pivote para μ basado en $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ y $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Es fácil ver que el pivote satisface las dos condiciones enunciadas al principio de la Sección 3. De acuerdo con los resultados expuestos en la sección 3.1

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} < t_{n-1, \alpha/2} \right\} + \mathbf{1} \left\{ \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} > t_{n-1, 1-\alpha/2} \right\}, \quad (30)$$

es un test de nivel α para decidir entre las hipótesis $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$. Dicho en palabras, el test en rechazar H_0 si $\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} < t_{n-1, \alpha/2}$ o $\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} > t_{n-1, 1-\alpha/2}$ y aceptarla en otro caso.

Ejemplo

Ejemplo 4.4. En la siguiente tabla se muestran las mediciones, en segundos de grado, obtenidas por James Short (1761), de la paralaje solar (ángulo bajo el que se ve el radio ecuatorial de la tierra desde el centro del sol) .

8.50	8.50	7.33	8.64	9.27	9.06	9.25	9.09	8.50	8.06
8.43	8.44	8.14	7.68	10.34	8.07	8.36	9.71	8.65	8.35
8.71	8.31	8.36	8.58	7.80	7.71	8.30	9.71	8.50	8.28
9.87	8.86	5.76	8.44	8.23	8.50	8.80	8.40	8.82	9.02
10.57	9.11	8.66	8.34	8.60	7.99	8.58	8.34	9.64	8.34
8.55	9.54	9.07							

Con esos datos tenemos que $\bar{X} = 8.6162$ y $S = 0.749$. En la Figura 5 se muestra un histograma de los datos.

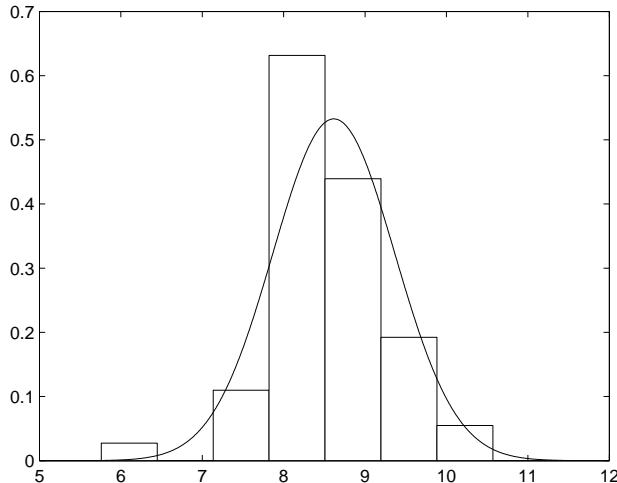


Figura 5: Histograma de las mediciones obtenidas por James Short. Parece razonable asumir que las mediciones de la paralaje solar tienen distribución normal.

Asumiendo que las mediciones tienen distribución $\mathcal{N}(\mu, \sigma^2)$ queremos decidir, con un nivel de significación $\alpha = 0.05$, entre las hipótesis

$$H_0 : \mu = 8.798 \quad \text{contra} \quad H_1 : \mu \neq 8.798$$

Como $n = 53$ y $t_{52, 0.025} = -t_{52, 0.975} = -2.0066$, el test de hipótesis (30) adopta la forma

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ \frac{\sqrt{53}(\bar{X} - 8.798)}{S} < -2.0066 \right\} + \mathbf{1} \left\{ \frac{\sqrt{53}(\bar{X} - 8.798)}{S} > 2.0066 \right\}.$$

Usando los datos de las mediciones tenemos que

$$\frac{\sqrt{53}(\bar{X} - 8.798)}{S} = \frac{\sqrt{53}(8.6162 - 8.798)}{0.749} = -1.7667.$$

Por lo tanto, no hay evidencia suficiente para rechazar que la paralaje solar es $\mu = 8.798$.

Usando como paralaje solar el valor $\mu = 8.798''$ y como radio ecuatorial de la tierra el valor $R = 6378$ km., trigonometría mediante, se puede determinar la distancia D entre la tierra y el sol:

$$\tan\left(\frac{8.798}{3600} \times \frac{\pi}{180}\right) = \frac{6378}{D} \iff D = 1.4953 \times 10^8.$$

Lo que significa que la distancia entre la tierra y el sol es 149.53 millones de km. \square

5. Test para probabilidad de éxito de distribuciones Bernoulli

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria con distribución Bernoulli(p), $p \in (0, 1)$. Basados en la muestra aleatoria, \mathbf{X} , queremos construir test para decidir entre dos hipótesis sobre la probabilidad de éxito p .

La cantidad de éxitos en la muestra

$$N = \sum_{i=1}^n X_i$$

tiene distribución Binomial(n, p) y resume toda la información relevante sobre el parámetro p contenida en la muestra aleatoria \mathbf{X} . La media y la varianza de N son, respectivamente, $\mathbb{E}_p[N] = np$ y $\mathbb{V}_p(N) = np(1 - p)$.

Lema 5.1 (Dominación estocástica). Sean $0 < p_1 < p_2 < 1$ arbitrarios pero fijos. Si $N_1 \sim \text{Binomial}(n, p_1)$ y $N_2 \sim \text{Binomial}(n, p_2)$, entonces para cada $x \in \mathbb{R}$ vale que

$$\mathbb{P}(N_2 \leq x) \leq \mathbb{P}(N_1 \leq x).$$

Demostración Sean U_1, \dots, U_n variables aleatorias independientes cada una con distribución $\mathcal{U}(0, 1)$. Para cada $i = 1, \dots, n$ construya las siguientes variables

$$X_{1,i} := \mathbf{1}\{U_i \leq p_1\}, \quad X_{2,i} := \mathbf{1}\{U_i \leq p_2\}.$$

Por construcción valen las siguientes propiedades:

- (a) las variables $X_{1,1}, \dots, X_{1,n}$ son iid Bernoulli(p_1);
- (b) las variables $X_{2,1}, \dots, X_{2,n}$ son iid Bernoulli(p_2);

(c) para cada i vale que $X_{2,i} \geq X_{1,i}$.

En consecuencia, las variables

$$\hat{N}_1 := \sum_{i=1}^n X_{1,i} \sim \text{Binomial}(n, p_1), \quad \hat{N}_2 := \sum_{i=1}^n X_{2,i} \sim \text{Binomial}(n, p_2)$$

verifican que $\hat{N}_1 \leq \hat{N}_2$. Se deduce entonces que $\{\hat{N}_2 \leq x\} \subseteq \{\hat{N}_1 \leq x\}$, para cualquier $x \in \mathbb{R}$. Por lo tanto,

$$\mathbb{P}(N_2 \leq x) = \mathbb{P}(\hat{N}_2 \leq x) \leq \mathbb{P}(\hat{N}_1 \leq x) = \mathbb{P}(N_1 \leq x).$$

□

Corolario 5.2. Sea N una variable aleatoria con distribución $\text{Binomial}(n, p)$, $p \in (0, 1)$. Fijado un valor $x \in \mathbb{R}^+$, la función polinómica de grado n , $h : (0, 1) \rightarrow [0, 1]$, definida por

$$h(p) = \mathbb{P}_p(N \leq x) = \sum_{k=0}^{[x]} \binom{n}{k} p^k (1-p)^{n-k}$$

es decreciente.

□

5.1. Test para moneda honesta (de lo simple a lo complejo)

Se quiere decidir si una moneda es honesta o no lo es. Formalmente, se trata de construir un test para decidir entre las hipótesis

$$H_0 : p = \frac{1}{2} \quad \text{contra} \quad H_1 : p \neq \frac{1}{2}.$$

1.- Se quiere decidir tirando la moneda 6 veces. ¿Qué hacer? Observamos la cantidad N de caras obtenidas en los 6 tiros. Para cada p tenemos que $N \sim \text{Binomial}(6, p)$. Cuando la moneda es honesta, $\mathbb{E}_{1/2}[N] = 3$. Teniendo en cuenta la existencia de fluctuaciones parece razonable aceptar que la moneda es honesta cuando observamos que $2 \leq N \leq 4$. Proponemos entonces el siguiente test

$$\delta(\mathbf{X}) = 1 - \mathbf{1}\{2 \leq N \leq 4\} = \mathbf{1}\{N < 2\} + \mathbf{1}\{N > 4\},$$

cuya función de potencia des

$$\beta(p) = \mathbb{P}_p(N \leq 1) + \mathbb{P}_p(N \geq 5) = (1-p)^6 + 6p(1-p)^5 + 6p^5(1-p) + p^6.$$

Dada una moneda honesta, ¿qué riesgo se corre de rechazarla como falsa? Esta pregunta se contesta calculando el nivel de significación del test $\alpha = \beta(1/2) = \frac{14}{64} = 0.21875$. □

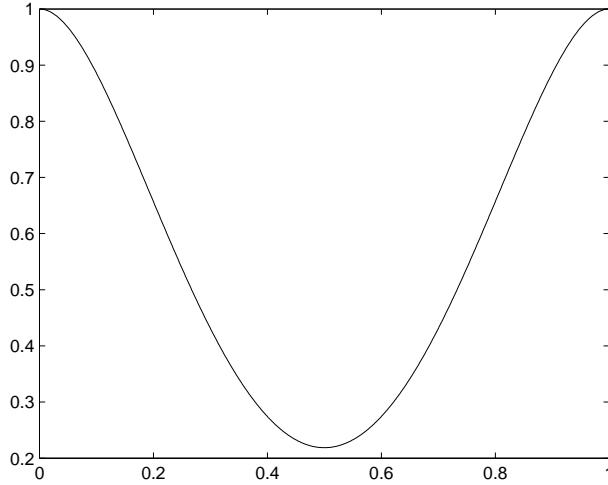


Figura 6: Gráfico de la función de potencia del test $\delta(\mathbf{X}) = \mathbf{1}\{N < 2\} + \mathbf{1}\{N > 4\}$.

2.- Se propone el siguiente *test*: lanzar la moneda 100 veces y contar la cantidad de caras observadas N . Si $40 \leq N \leq 60$ se decide que la moneda es honesta. En caso contrario, se decide que no lo es.

Definido el test lo único que queda por hacer es evaluar los riesgos de decisiones erróneas. Para ello calculamos la función de potencia

$$\beta(p) = \mathbb{P}(\text{Rechazar } H_0 | p) = \mathbb{P}_p(N < 40) + \mathbb{P}_p(N > 60).$$

Para cada p la cantidad de caras observadas en 100 lanzamientos se distribuye como una Binomial: $N \sim \text{Binomial}(100, p)$. En consecuencia,

$$\beta(p) = \sum_{k=0}^{39} \binom{100}{k} p^k (1-p)^{100-k} + \sum_{k=61}^{100} \binom{100}{k} p^k (1-p)^{100-k}. \quad (31)$$

Sin una herramienta computacional a la mano es insensato calcular riesgos utilizando la expresión obtenida en (31). Como el volumen de la muestra es 100 usando el teorema central del límite, $N \sim \mathcal{N}(100p, 100p(1-p))$, podemos obtener una buena aproximación de la función de potencia, (al menos para valores de p contenidos en el intervalo abierto $(0.12, 0.88)$)

$$\begin{aligned} \beta(p) &\approx \Phi\left(\frac{40 - 100p}{\sqrt{100p(1-p)}}\right) + 1 - \Phi\left(\frac{60 - 100p}{\sqrt{100p(1-p)}}\right) \\ &= \Phi\left(\frac{4 - 10p}{\sqrt{p(1-p)}}\right) + \Phi\left(\frac{10p - 6}{\sqrt{p(1-p)}}\right) \end{aligned} \quad (32)$$

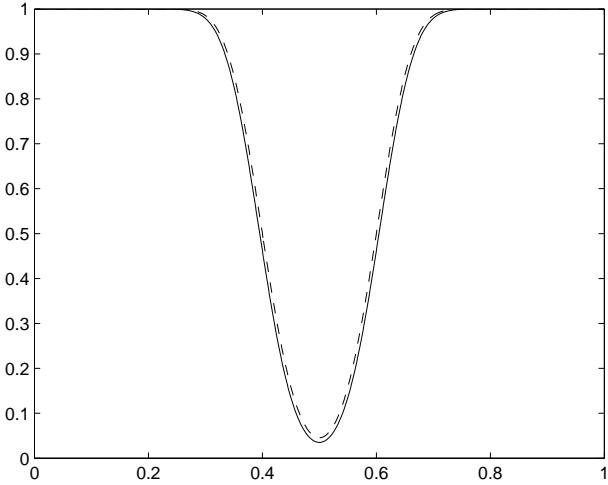


Figura 7: Gráfico de la función de potencia del test $\delta(\mathbf{X}) = \mathbf{1}\{N < 40\} + \mathbf{1}\{N > 60\}$. En línea quebrada aproximación usando el TCL.

Es más o menos claro que la función de potencia es simétrica respecto de $p = 1/2$. Esto es, para cada $q \in (0, 1/2)$, vale que $\beta(1/2 - q) = \beta(1/2 + q)$.

Riesgos:

1. El *nivel de significación del test* es $\alpha = \beta(1/2)$. Calculamos $\beta(1/2)$ utilizando la aproximación obtenida en (32)

$$\beta(1/2) \approx \Phi\left(\frac{4 - 5}{\sqrt{1/4}}\right) + \Phi\left(\frac{5 - 6}{\sqrt{1/4}}\right) = \Phi(-2) + \Phi(-2) \approx 0.0455$$

Esto significa que la probabilidad de rechazar que la moneda es honesta, cuando en verdad lo es, será 0.0455. En palabras: de cada 100 monedas honestas sometidas a verificación (en promedio) serán rechazadas como falsas 4 o 5 de ellas.

2. ¿Qué riesgo se corre de aceptar como honesta una moneda falsa, con carga 0.7 hacia el lado de la cara? Para contestar esta pregunta tenemos que calcular el valor de $1 - \beta(0.7)$. Usando (32) obtenemos

$$1 - \beta(0.7) \approx 1 - \Phi\left(\frac{4 - 7}{\sqrt{0.21}}\right) - \Phi\left(\frac{7 - 6}{\sqrt{0.21}}\right) \approx 0.0146.$$

Grosso modo el resultado se interpreta de la siguiente manera: de cada 100 monedas cargadas con 0.7 para el lado de cara sometidas a verificación (en promedio) serán aceptadas como honestas 1 o 2 de ellas. \square

3.- Queremos un test de nivel de significación $\alpha = 0.05$, basado en 64 lanzamientos de la moneda. Parece razonable proponer un test de la forma

$$\delta(\mathbf{X}) = \mathbf{1}\{N < 32 - k\} + \mathbf{1}\{N > 32 + k\}.$$

El problema consiste en determinar el valor de k . El nivel de significación del test es

$$\beta(1/2) = \mathbb{P}_{1/2}(N < 32 - k) + \mathbb{P}_{1/2}(N > 32 + k)$$

Para $p = 1/2$, $N \sim \text{Binomial}(64, 1/2)$ y usando el teorema central de límite obtenemos que la distribución de N es aproximadamente normal de media $\mathbb{E}_{1/2}[N] = (1/2)64 = 32$ y varianza $V_{1/2}(N) = (1/2)(1/2)64 = 16$.

$$\begin{aligned} \beta(1/2) &= \mathbb{P}_{1/2}(N < 32 - k) + \mathbb{P}_{1/2}(N > 32 + k) \\ &\approx \mathbb{P}_{1/2}\left(\frac{N - 32}{4} < -\frac{k}{4}\right) + \mathbb{P}_{1/2}\left(\frac{N - 32}{4} > \frac{k}{4}\right) \\ &= \Phi\left(-\frac{k}{4}\right) + \Phi\left(-\frac{k}{4}\right) = 2\Phi\left(-\frac{k}{4}\right) \end{aligned}$$

En consecuencia,

$$\beta(1/2) = 0.05 \iff \Phi\left(-\frac{k}{4}\right) = 0.025 \iff -\frac{k}{4} = z_{0.025} = -1.96 \iff k = 7.84.$$

Por lo tanto, el test adopta la forma

$$\delta(\mathbf{X}) = \mathbf{1}\{N < 32 - 7.84\} + \mathbf{1}\{N > 32 + 7.84\} = \mathbf{1}\{N < 25\} + \mathbf{1}\{N > 39\}.$$

En palabras, el *test* consiste en lo siguiente: lanzar la moneda 64 veces; si la cantidad de caras observadas es menor que 25 o mayor que 39, se decide que la moneda está cargada; en caso contrario, se decide que la moneda es honesta.

¿Qué riesgo se corre de aceptar como honesta una moneda con carga 0.7 hacia el lado de la cara? La respuesta se obtiene calculando $1 - \beta(0.7)$. Para $p = 0.7$ el TCL establece que $(N - 0.7(64))/\sqrt{(0.7)(0.3)64} \sim \mathcal{N}(0, 1)$, en consecuencia,

$$\beta(0.7) \approx \Phi\left(\frac{25 - 0.7(64)}{\sqrt{(0.21)64}}\right) + \Phi\left(\frac{0.7(64) - 39}{\sqrt{(0.21)64}}\right) \approx \Phi(1.5821) = 0.94318.$$

Por lo tanto, $1 - \beta(0.7) = 0.0568\dots$

□

4.- Queremos un test de nivel de significación $\alpha = 0.05$, cuya potencia cuando la carga difiere de 0.5 en más de 0.1 sea como mínimo 0.90. Parece razonable proponer una regla de la forma

$$\delta(\mathbf{X}) = \mathbf{1}\{N < n(1/2) - k\} + \mathbf{1}\{N > n(1/2) + k\}.$$

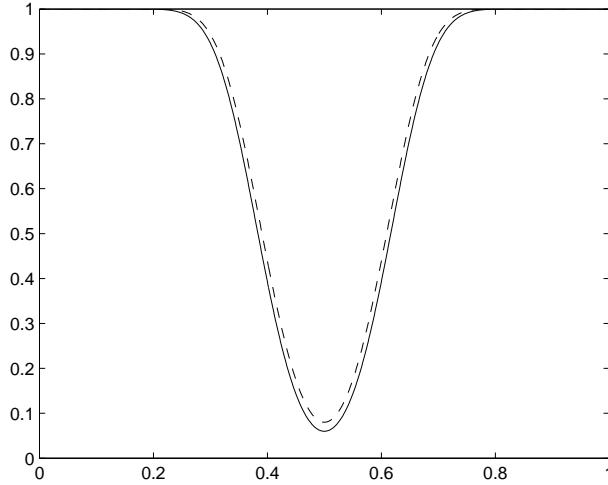


Figura 8: Gráfico de la función de potencia del test $\delta(\mathbf{X}) = \mathbf{1}\{N < 25\} + \mathbf{1}\{N > 39\}$. En línea quebrada aproximación usando el TCL.

El problema consiste en determinar el volumen de la muestra, n , y el valor de k . Las condiciones impuestas al test pueden expresarse de la siguiente manera

$$\alpha(\delta) \leq 0.05 \quad \text{y} \quad \beta(0.6) \geq 0.90, \quad (33)$$

donde $\alpha(\delta) = \beta(1/2)$ es en nivel del test y $\beta(0.6)$ es la potencia en $p = 0.6$.

Ambos problemas se resuelven caracterizando la función de potencia del test

$$\beta(p) = \mathbb{P}_p(N < n(1/2) - n\epsilon) + \mathbb{P}_p(N > n(1/2) + n\epsilon)$$

De acuerdo con el el TCL tenemos que para cada p

$$Z = \frac{N - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0, 1),$$

en consecuencia,

$$\begin{aligned} \beta(p) &\approx \mathbb{P}_p\left(Z < \frac{n(1/2 - p) - n\epsilon}{\sqrt{np(1-p)}}\right) + \mathbb{P}_p\left(Z > \frac{n(1/2 - p) + n\epsilon}{\sqrt{np(1-p)}}\right) \\ &= \Phi\left(\frac{\sqrt{n}(1/2 - p - \epsilon)}{\sqrt{p(1-p)}}\right) + \Phi\left(\frac{\sqrt{n}(p - 1/2 - \epsilon)}{\sqrt{p(1-p)}}\right) \end{aligned}$$

Notar que para $p > 1/2$ el primer término del lado derecho de la igualdad es despreciable y entonces

$$\beta(0.6) \approx \Phi\left(\frac{\sqrt{n}(0.1 - \epsilon)}{\sqrt{0.24}}\right)$$

Por otra parte,

$$\beta(1/2) \approx 2\Phi\left(\frac{-\sqrt{n}\epsilon}{\sqrt{1/4}}\right) = 2\Phi(-2\sqrt{n}\epsilon)$$

En consecuencia, las desigualdades (33) son equivalentes a las siguientes:

$$2\Phi(-2\sqrt{n}\epsilon) \leq 0.05 \quad \text{y} \quad \Phi\left(\frac{\sqrt{n}(0.1 - \epsilon)}{\sqrt{0.24}}\right) \geq 0.90.$$

Por lo tanto, n y ϵ deben ser tales que

$$2\epsilon\sqrt{n} \geq z_{0.975} \quad \text{y} \quad \frac{\sqrt{n}(0.1 - \epsilon)}{\sqrt{0.24}} \geq z_{0.90} \quad (34)$$

Recurriendo a una tabla de la distribución normal, usando una calculadora de almacenero (que tenga una tecla con el símbolo $\sqrt{\cdot}$), y operando con las desigualdades (34) se pueden obtener soluciones particulares. Por ejemplo, $n = 259$ y $\epsilon = 0.061$.

Tomando $n = 259$ y $\epsilon = 0.061$ obtenemos la siguiente regla de decisión:

$$\delta(\mathbf{X}) = \mathbf{1}\{N < 114\} + \mathbf{1}\{N > 145\}.$$

En palabras, el test establece que hay que lanzar la moneda 259 veces y contar la cantidad de caras observadas. Si la cantidad de caras observadas es menor que 114 o mayor que 145 se decide que la moneda está cargada. En caso contrario, se decide que es honesta.

Una cuenta. Para obtener el resultado particular $n = 259$ y $\epsilon = 0.061$ hay que hacer lo siguiente: En primer lugar, hay que observar que

$$\begin{aligned} \frac{\sqrt{n}(0.1 - \epsilon)}{\sqrt{0.24}} \geq z_{0.90} &\iff \sqrt{n}(0.1 - \epsilon) \geq z_{0.90}\sqrt{0.24} \\ &\iff 0.1\sqrt{n} - z_{0.90}\sqrt{0.24} \geq \epsilon\sqrt{n} \\ &\iff 2(0.1\sqrt{n} - z_{0.90}\sqrt{0.24}) \geq 2\epsilon\sqrt{n} \end{aligned} \quad (35)$$

La última desigualdad de (35) combinada con la primera de (34) implican que n debe satisfacer las desigualdades

$$\begin{aligned} 0.2\sqrt{n} - 2z_{0.90}\sqrt{0.24} \geq z_{0.975} &\iff \sqrt{n} \geq 5(z_{0.975} + 2z_{0.90}\sqrt{0.24}) \\ &\iff n \geq 25(z_{0.975} + 2z_{0.90}\sqrt{0.24})^2 \end{aligned}$$

Tabla de la distribución normal ($z_{0.975} = 1.96$, $z_{0.90} = 1.28$) y calculadora mediante, se obtiene que $n \geq 259$. Poniendo $n = 259$ en la tercera desigualdad de (35) se puede ver que ϵ debe ser tal que

$$\epsilon \leq 0.1 - z_{0.90} \frac{\sqrt{0.24}}{\sqrt{259}} \approx 0.061.$$

Podemos elegir $\epsilon = 0.061$. □

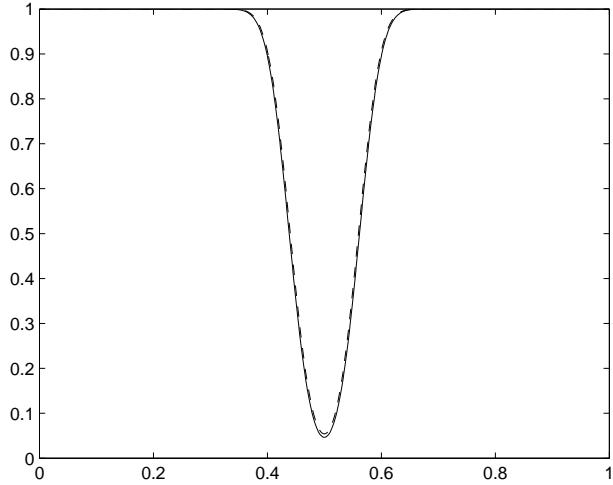


Figura 9: Gráfico de la función de potencia del test $\delta(\mathbf{X}) = \mathbf{1}\{N < 114\} + \mathbf{1}\{N > 145\}$. En línea quebrada aproximación usando el TCL.

5.2. Hipótesis fundamental simple

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria con distribución Bernoulli(p), $p \in (0, 1)$. Basados en la muestra aleatoria \mathbf{X} queremos construir test para decidir entre las hipótesis

$$H_0 : p = p_0 \quad \text{contra} \quad H_1 : p \neq p_0,$$

donde $p_0 \in (0, 1)$ es un valor arbitrario pero fijo.

Primera fase: diseñar un test de hipótesis

Cuando la hipótesis H_0 es verdadera, la cantidad de éxitos $N = \sum_{i=1}^n X_i$ tiene distribución binomial de media np_0 y desvío $\sqrt{np_0(1 - p_0)}$. Parece razonable construir reglas de decisión de la forma

$$\delta(\mathbf{X}) = \mathbf{1}\{N < np_0 - n\epsilon\} + \mathbf{1}\{N > np_0 + n\epsilon\}, \quad (36)$$

donde $n \in \mathbb{N}$ y $\epsilon > 0$ son arbitrarios pero fijos.

En castellano, el test de hipótesis definido en (36) establece el siguiente procedimiento de decisión:

1. Examinar una muestra de tamaño n de la variable aleatoria Bernoulli, $\mathbf{X} = (X_1, \dots, X_n)$ y contar la cantidad de éxitos observados: $N = \sum_{i=1}^n X_i$.

2. Si la cantidad de éxitos observados es menor que $np_0 - n\epsilon$ o mayor que $np_0 + n\epsilon$ se rechaza la hipótesis $p = p_0$ y se decide que $p \neq p_0$. En caso contrario, se no se rechaza la hipótesis $p = p_0$.

Segunda fase: caracterizar la función de potencia

La *segunda fase* del programa consiste en “calcular” la función de potencia. Esta función permite calcular los riesgos de tomar decisiones erróneas:

$$\begin{aligned}\beta(p) &= \mathbb{P}(\text{Rechazar } H_0 | p) = \mathbb{P}_p(\delta(\mathbf{X}) = 1) \\ &= \mathbb{P}_p(N < np_0 - n\epsilon) + \mathbb{P}_p(N > np_0 + n\epsilon) \\ &= \sum_{k=0}^{[np_0 - n\epsilon]} \binom{n}{k} p^k (1-p)^{n-k} + \sum_{k=[np_0 - n\epsilon] + 1}^n \binom{n}{k} p^k (1-p)^{n-k}. \quad (37)\end{aligned}$$

Notar que la función de potencia resultó ser un complicado polinomio de grado n y no es fácil capturar a simple vista su comportamiento cualitativo.

Nivel de significación. Debido a que la hipótesis fundamental es de la forma $p = p_0$, para cada n y ϵ , el *nivel de significación del test* es

$$\alpha(\delta) = \beta(p_0) = \sum_{k=0}^{[np_0 - n\epsilon]} \binom{n}{k} p_0^k (1-p_0)^{n-k} + \sum_{k=[np_0 - n\epsilon] + 1}^n \binom{n}{k} p_0^k (1-p_0)^{n-k}. \quad (38)$$

Nota Bene 1. Notar que los test (36) contienen un juego de dos parámetros, n y ϵ . Estos parámetros determinan la calidad de cada test y deben ajustarse de acuerdo con las prescripciones impuestas al test sobre su nivel de significación y su potencia en alguna hipótesis alternativa.

Nota Bene 2. Notar que si la muestra tiene volumen prefijado n , por más que se mueva el valor de ϵ , el nivel de significación del test $\alpha(\delta)$ puede tomar a lo sumo $n + 1$ valores distintos. Por lo tanto, si se prescribe que el nivel de significación del test $\delta(\mathbf{X})$ debe ser α , casi seguramente la ecuación $\alpha(\delta) = \alpha$ no tendrá solución.

Aproximación por TCL para muestras “grandes”

La función de potencia (37) se puede aproximar utilizando el teorema central del límite. Si la muestra es suficientemente grande, para cada valor de p , tenemos que

$$Z = \frac{N - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0, 1).$$

Esto permite aproximar el valor de $\beta(p)$ de la siguiente manera

$$\begin{aligned}\beta(p) &= \mathbb{P}_p \left(Z < \frac{n(p_0 - p - \epsilon)}{\sqrt{np(1-p)}} \right) + \mathbb{P}_p \left(Z > \frac{n(p_0 - p + \epsilon)}{\sqrt{np(1-p)}} \right) \\ &\approx \Phi \left(\frac{\sqrt{n}(p_0 - p - \epsilon)}{\sqrt{p(1-p)}} \right) + \Phi \left(\frac{\sqrt{n}(p - p_0 - \epsilon)}{\sqrt{p(1-p)}} \right).\end{aligned}\quad (39)$$

Aunque la aproximación (39) pueda resultar “grosera” y no sea lo suficientemente buena para todos los posibles valores de p , permite capturar el comportamiento cualitativo de la función de potencia.

Nivel de significación. Poniendo $p = p_0$, la aproximación (39) permite observar que

$$\alpha(\delta) = \beta(p_0) = 2\Phi \left(\frac{-\sqrt{n}\epsilon}{\sqrt{p_0(1-p_0)}} \right). \quad (40)$$

Esto indica que basta tomar n suficientemente grande para que $\beta(p_0)$ se ubique todo lo cerca del 0 que uno quiera. En otras palabras, el test puede construirse para garantizar que la probabilidad de rechazar la hipótesis $p = p_0$ cuando ella es verdadera sea todo lo chica que uno quiera.

La aproximación (40) se puede utilizar para ajustar los valores de los parámetros n y ϵ para que valga la desigualdad $\alpha(\delta) \leq \alpha$. Para ello basta observar que la desigualdad aproximada

$$2\Phi \left(\frac{-\sqrt{n}\epsilon}{\sqrt{p_0(1-p_0)}} \right) \leq \alpha \iff \frac{-\sqrt{n}\epsilon}{\sqrt{p_0(1-p_0)}} \leq z_{\alpha/2}. \quad (41)$$

Por lo tanto, las soluciones de la desigualdad (41) serán todos los valores de $n \in \mathbb{N}$ y todos los valores de $\epsilon > 0$ que satisfagan

$$\frac{\sqrt{n}\epsilon}{\sqrt{p_0(1-p_0)}} \geq z_{1-\alpha/2}. \quad (42)$$

Fijada una solución particular de (42), una alta dosis de paciencia permite calcular a mano el valor exacto del nivel de significación $\alpha(\delta)$ obtenido en (38) y comprobar si efectivamente satisface $\alpha(\delta) \leq \alpha$.

Test de hipótesis con nivel de significación aproximado. Basados en los argumentos y razonamientos anteriores, podemos diseñar test para decidir entre las hipótesis $H_0 : p = p_0$ contra $H_1 : p \neq p_0$ con nivel de significación “aproximadamente” α . Usando el diseño (36) para valores de n y ϵ que verifiquen la desigualdad (42) obtenemos

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ N < np_0 - z_{1-\alpha/2} \sqrt{np_0(1-p_0)} \right\} + \mathbf{1} \left\{ N > np_0 + z_{1-\alpha/2} \sqrt{np_0(1-p_0)} \right\}. \quad (43)$$

Potencia en una alternativa. El mismo problema se presenta cuando se prescribe una potencia β para una alternativa p_1 . En esta situación trataremos de resolver la desigualdad $\beta(p_1) \geq \beta$. Nuevamente la aproximación (39) permite resolver el problema:

- Si $p_1 < p_0$ el segundo término en (39) es despreciable respecto del primero y entonces obtenemos la siguiente aproximación:

$$\beta(p_1) \approx \Phi \left(\frac{\sqrt{n}(p_0 - p_1 - \epsilon)}{\sqrt{p_1(1 - p_1)}} \right). \quad (44)$$

- Si $p_1 > p_0$ el primer término es despreciable respecto del segundo y entonces obtenemos la siguiente aproximación:

$$\beta(p_1) \approx \Phi \left(\frac{\sqrt{n}(p_1 - p_0 - \epsilon)}{\sqrt{p_1(1 - p_1)}} \right). \quad (45)$$

Para fijar ideas supongamos que $p_1 > p_0$. Razonando del mismo modo que antes se obtiene la siguiente solución “aproximada” de la inecuación $\beta(p_1) \geq \beta$:

$$\frac{\sqrt{n}(p_1 - p_0 - \epsilon)}{\sqrt{p_1(1 - p_1)}} \geq z_\beta. \quad (46)$$

El razonamiento anterior muestra que, prefijados dos valores α y β , se pueden diseñar test de hipótesis de la forma (36) con prescripciones del siguiente tipo: nivel de significación menor o igual que α y/o potencia en una alternativa particular superior a β . \square

5.3. Hipótesis fundamental compuesta

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una variable aleatoria con distribución Bernoulli(p), $p \in (0, 1)$. Basados en la muestra aleatoria \mathbf{X} queremos construir test para decidir entre las hipótesis

$$H_0 : p \leq p_0 \quad \text{contra} \quad H_1 : p > p_0,$$

donde $p_0 \in (0, 1)$ es un valor arbitrario pero fijo.

Programa de actividades. Adaptaremos los argumentos y razonamientos desarrollados en la sección 5.2. La primera fase del programa consiste en construir test de hipótesis basados en la cantidad de éxitos de la muestra $N = \sum_{i=1}^n X_i$. La segunda fase del programa consiste en evaluar los riesgos de tomar decisiones erróneas con los test construidas: se trata de caracterizar analíticamente la función de potencia y estudiar sus propiedades cualitativas y cuantitativas: cálculo del nivel de significación y de la potencia en las hipótesis alternativas simples.

Test de hipótesis. En este caso resulta intuitivamente claro proponer test de forma

$$\delta(\mathbf{X}) = \mathbf{1}\{N > np_0 + n\epsilon\}, \quad (47)$$

donde n y ϵ son parámetros ajustables.

Función de potencia. Fijados n y ϵ la función de potencia del test es

$$\begin{aligned} \beta(p) &= \mathbb{P}(\text{rechazar } H_0 | p) = \mathbb{P}_p(\delta(\mathbf{X}) = 1) = \mathbb{P}_p(N > np_0 + n\epsilon) \\ &= \sum_{k=[np_0+n\epsilon]+1}^n \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned} \quad (48)$$

De acuerdo con el **Corolario 5.2** la función de potencia es creciente. Esto es intuitivamente claro si se piensa que cuando aumenta la probabilidad de cada éxito, la cantidad de éxitos debe aumentar.

Aproximación por TCL. Si el volumen de muestra es suficientemente grande, usando el teorema central del límite podemos obtener la siguiente expresión aproximada de la función de potencia

$$\beta(p) = \mathbb{P}_p \left(\frac{N - np}{\sqrt{np(1-p)}} > \frac{np_0 + n\epsilon - np}{\sqrt{np(1-p)}} \right) \approx \Phi \left(\frac{\sqrt{n}(p - p_0 - \epsilon)}{\sqrt{p(1-p)}} \right). \quad (49)$$

Nivel de significación. Como la función de potencia es creciente, el nivel de significación del test se obtiene de la siguiente manera

$$\alpha(\delta) = \max_{p \leq p_0} \beta(p) = \beta(p_0) = \sum_{k=[np_0+n\epsilon]+1}^n \binom{n}{k} p_0^k (1-p_0)^{n-k} \approx \Phi \left(\frac{-\sqrt{n}\epsilon}{\sqrt{p_0(1-p_0)}} \right). \quad (50)$$

La aproximación en (50) presupone que el volumen de muestra es suficientemente grande (por ejemplo, $np_0(1-p_0) > 10$).

Prefijados un volumen de muestra suficientemente grande y un nivel de significación α para el test de hipótesis, la aproximación (50) permite hallar el valor de ϵ

$$z_{1-\alpha} \sqrt{p_0(1-p_0)} = \sqrt{n}\epsilon. \quad (51)$$

Test de hipótesis con nivel de significación aproximado. Usando el diseño (47) y el resultado obtenido en (51) se deduce que, para n suficientemente grande y fijo, la forma del test de hipótesis de nivel de significación α para decidir entre $H_0 : p \leq p_0$ contra $H_1 : p > p_0$ es

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ N > np_0 + z_{1-\alpha} \sqrt{np_0(1-p_0)} \right\}. \quad (52)$$

Potencia en una alternativa. El análisis de la potencia en las hipótesis alternativas simples $p = p_1$, con $p_1 > p_0$, se realiza siguiendo las mismas líneas desarrolladas en la sección anterior. \square

Ejemplo 5.3. Un productor de chips afirma que no más del 2% de los chips que produce son defectuosos. Una compañía electrónica (impresionada por dicha afirmación) le compra una gran cantidad de chips. Para determinar si la afirmación del productor se puede tomar literalmente, la compañía decide testear una muestra de 300 de esos chips. Si se encuentra que 10 de los 300 chips son defectuosos, debería rechazarse la afirmación del productor?

Solución. Formalmente, el problema consiste en construir un test de hipótesis para decidir entre

$$H_0 : p \leq 0.02 \quad \text{contra} \quad H_1 : p > 0.02.$$

sobre la base de una muestra de volumen 300.

Fijado un nivel de significación, por ejemplo $\alpha = 0.05$, el test de hipótesis (52) adopta la forma

$$\begin{aligned} \delta(\mathbf{X}) &= \mathbf{1} \left\{ N > 300(0.02) + z_{0.95} \sqrt{300(0.02)(0.98)} \right\} = \mathbf{1}\{N > 9.9886\} \\ &= \mathbf{1}\{N \geq 10\}. \end{aligned} \tag{53}$$

Dicho en palabras, al nivel del 5% de significación, un test para decidir entre las hipótesis $H_0 : p \leq 0.02$ contra $H_1 : p > 0.02$, basado en una muestra de volumen 300, consiste en rechazar la hipótesis H_0 siempre que se observen 10 o más éxitos.

Traducido al problema que estamos examinando, el criterio de decisión puede enunciarse de la siguiente manera: “examinar 300 componentes. Si se observan 10 o más defectuosos debe rechazarse la afirmación del productor de que produce con una calidad de a lo sumo un 2%, si se observan menos de 10 defectuosos no hay evidencia suficiente para rechazar su afirmación.”

En conclusión, como en la muestra examinada se observaron 10 chips defectuosos, al nivel del 5% de significación, la afirmación del productor debe rechazarse. \square

6. Test para varianza de normales

El objetivo de esta sección es ilustrar cómo se pueden obtener test de hipótesis usando intervalos de confianza.

6.1. Hipótesis sobre varianza con media conocida

Usando intervalos de confianza para la varianza de una distribución normal $\mathcal{N}(\mu, \sigma^2)$ con media μ conocida vamos a construir test de hipótesis de nivel de significación α para decidir entre

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{contra} \quad H_1 : \sigma^2 \neq \sigma_0^2,$$

para algún valor σ_0^2 determinado.

Dada una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de la distribución normal $\mathcal{N}(\mu, \sigma^2)$ con media μ conocida, sabemos que

$$I(\mathbf{X}) = \left[\frac{n\widehat{\sigma}_{mv}^2}{\chi_{n,(1+\beta)/2}^2}, \frac{n\widehat{\sigma}_{mv}^2}{\chi_{n,(1-\beta)/2}^2} \right],$$

donde $n\widehat{\sigma}_{mv}^2 = \sum_{i=1}^n (X_i - \mu)^2$, es un intervalo de confianza para σ^2 de nivel β . Poniendo $\beta = 1 - \alpha$ se obtiene el siguiente test de nivel α para decidir entre las hipótesis $H_0 : \sigma^2 = \sigma_0^2$ contra $H_1 : \sigma^2 \neq \sigma_0^2$

$$\begin{aligned} \delta(\mathbf{X}) &= \mathbf{1}\{I(\mathbf{X}) \not\ni \sigma_0^2\} \\ &= \mathbf{1}\left\{ \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \mu)^2 < \chi_{n,\alpha/2}^2 \right\} + \mathbf{1}\left\{ \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \mu)^2 > \chi_{n,1-\alpha/2}^2 \right\}. \end{aligned} \quad (54)$$

Función de potencia. Para calcular y analizar el comportamiento de la función de potencia,

$$\beta(\sigma^2) = \mathbb{P}(\text{Rechazar } H_0 | \sigma^2),$$

debe recordarse que cuando el verdadero valor de la varianza es σ^2 , la variable aleatoria $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$ tiene distribución $\chi_n^2 = \Gamma(n/2, 1/2)$. Multiplicando por $\frac{\sigma_0^2}{\sigma^2}$ en las desigualdades dentro de las llaves en la fórmula del test (54), y “calculando” las correspondientes probabilidades, obtenemos la siguiente expresión

$$\beta(\sigma^2) = \int_0^{a(\sigma^2)} \frac{(1/2)^{n/2}}{\Gamma(n/2)} x^{(n/2)-1} e^{-\frac{1}{2}x} dx + \int_{b(\sigma^2)}^{\infty} \frac{(1/2)^{n/2}}{\Gamma(n/2)} x^{(n/2)-1} e^{-\frac{1}{2}x} dx,$$

donde

$$a(\sigma^2) = \frac{\sigma_0^2}{\sigma^2} \chi_{n,\alpha/2}^2, \quad b(\sigma^2) = \frac{\sigma_0^2}{\sigma^2} \chi_{n,1-\alpha/2}^2.$$

□

Ejemplo 6.1. Dada una muestra aleatoria de volumen 10 de una población normal de media 0 se quiere construir un test de nivel $\alpha = 0.05$ para decidir entre las hipótesis $H_0 : \sigma^2 = 1$ contra $H_1 : \sigma^2 \neq 1$.

Solución. Como $\chi_{10,0.025}^2 = 3.247$ y $\chi_{10,0.975}^2 = 20.483$, el test de hipótesis (54) adopta la forma

$$\delta(\mathbf{X}) = \mathbf{1}\left\{ \sum_{i=1}^n X_i^2 < 3.247 \right\} + \mathbf{1}\left\{ \sum_{i=1}^n X_i^2 > 20.483 \right\}. \quad (55)$$

□

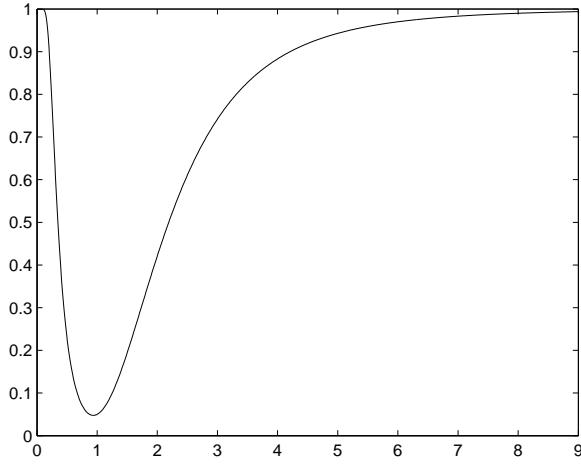


Figura 10: Gráfico de la función de potencia del test (55).

6.2. Hipótesis sobre varianza con media desconocida

Usando intervalos de confianza para la varianza de una distribución normal $\mathcal{N}(\mu, \sigma^2)$ vamos a construir test de hipótesis de nivel de significación α para decidir entre

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{contra} \quad H_1 : \sigma^2 \neq \sigma_0^2,$$

para algún valor σ_0^2 determinado.

Dada una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de la distribución normal $\mathcal{N}(\mu, \sigma^2)$ sabemos que

$$I(\mathbf{X}) = \left[\frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2} \right],$$

es un intervalo de confianza para σ^2 de nivel β . Poniendo $\beta = 1 - \alpha$ se obtiene el siguiente test de nivel α para decidir entre las hipótesis $H_0 : \sigma^2 = \sigma_0^2$ contra $H_1 : \sigma^2 \neq \sigma_0^2$

$$\begin{aligned} \delta(\mathbf{X}) &= \mathbf{1}\{I(\mathbf{X}) \not\ni \sigma_0^2\} \\ &= \mathbf{1}\left\{\frac{(n-1)S^2}{\sigma_0^2} < \chi_{n-1, \alpha/2}^2\right\} + \mathbf{1}\left\{\frac{(n-1)S^2}{\sigma_0^2} > \chi_{n-1, 1-\alpha/2}^2\right\}. \end{aligned} \quad (56)$$

Función de potencia. Notar que el análisis de función de potencia de test (56) es completamente análogo al desarrollado para el caso en que suponíamos que la media μ es conocida. \square

Nota Bene. Notar que los test de hipótesis definidas en (54) y (56) son inmediatamente útiles para tomar decisiones.

Ejemplo 6.2. En la Sección dedicada al estudio de intervalos de confianza mostramos que cuando una muestra aleatoria \mathbf{X} (de volumen 8) de una población normal $\mathcal{N}(\mu, \sigma^2)$ arroja los valores 9, 14, 10, 12, 7, 13, 11, 12, el intervalo $I_{\sigma^2} = [2.248, 21.304]$ es un intervalo de confianza de nivel $\beta = 0.95$ para la varianza σ^2 .

Si se quiere decidir al 5 % de significación entre las hipótesis

$$H_0 : \sigma^2 = 4 \quad \text{contra} \quad H_1 : \sigma^2 \neq 4.$$

el test de hipótesis (56) conduce a no rechazar la hipótesis $\sigma^2 = 4$. □

7. Comparación de dos muestras

7.1. Test para medias de dos muestras normales.

Sean $\mathbf{X} = (X_1, \dots, X_m)$ e $\mathbf{Y} = (Y_1, \dots, Y_n)$ dos muestras aleatorias independientes de distribuciones normales $\mathcal{N}(\mu_X, \sigma_X^2)$ y $\mathcal{N}(\mu_Y, \sigma_Y^2)$, respectivamente. Sea $\Delta = \mu_X - \mu_Y$. Queremos un test para decidir entre las hipótesis

$$H_0 : \Delta = 0 \quad \text{contra} \quad H_1 : \Delta > 0.$$

7.1.1. Varianzas conocidas

Supongamos que las varianzas σ_X^2 y σ_Y^2 son conocidas. Para construir el test de hipótesis usaremos los estimadores de media: \bar{X} y \bar{Y} . Puesto que

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\Delta, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)$$

el test de nivel α decidir entre $H_0 : \Delta = 0$ contra $H_1 : \Delta > 0$ es

$$\delta(\mathbf{X}, \mathbf{Y}) = \mathbf{1} \left\{ \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} > z_{1-\alpha} \right\}$$

□

7.1.2. Varianzas desconocidas pero iguales.

Supongamos las varianzas $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. En tal caso, bajo la hipótesis $\Delta = 0$ tenemos que

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2 \sqrt{\frac{1}{m} + \frac{1}{n}}}} \sim \mathcal{N}(0, 1).$$

Para estimar la varianza σ^2 ponderamos “adecuadamente” los estimadores de varianza S_X^2 y S_Y^2 ,

$$S_P^2 := \frac{m-1}{m+n-2} S_X^2 + \frac{n-1}{m+n-2} S_Y^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}.$$

Se puede mostrar que

$$U = \frac{(n+m-2)}{\sigma^2} S_P^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \sim \chi_{n+m-2}.$$

Debido a que las variables Z y U son independientes, tenemos que

$$T = \frac{Z}{\sqrt{U/(m+n-2)}} = \frac{\bar{X} - \bar{Y}}{\sqrt{S_P^2} \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

Por lo tanto,

$$\delta(\mathbf{X}, \mathbf{Y}) = \mathbf{1} \left\{ \frac{\bar{X} - \bar{Y}}{\sqrt{S_P^2} \sqrt{\frac{1}{m} + \frac{1}{n}}} > t_{m+n-2, 1-\alpha} \right\}.$$

es un test de nivel de significación α para decidir entre las hipótesis $H_0 : \Delta = 0$ contra $H_1 : \Delta > 0$. \square

7.2. Test F para varianzas de normales.

Sean $\mathbf{X} = (X_1, \dots, X_m)$ e $\mathbf{Y} = (Y_1, \dots, Y_n)$ dos muestras aleatorias independientes de distribuciones normales $\mathcal{N}(\mu_X, \sigma_X^2)$ y $\mathcal{N}(\mu_Y, \sigma_Y^2)$, respectivamente. Sea $R = \sigma_X^2/\sigma_Y^2$. Queremos un test para decidir entre las hipótesis

$$H_0 : R = 1 \quad \text{contra} \quad H_1 : R \neq 1.$$

Las varianzas σ_X^2 y σ_Y^2 se pueden estimar mediante sus estimadores insesgados S_X^2 y S_Y^2 . Las variables

$$U = \frac{(m-1)}{\sigma_X^2} S_X^2 \sim \chi_{m-1}^2 \quad \text{y} \quad V = \frac{(n-1)}{\sigma_Y^2} S_Y^2 \sim \chi_{n-1}^2$$

son independientes.

Test de hipótesis. Bajo la hipótesis $H_0 : R = 1$, vale que

$$F = \frac{S_X^2}{S_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{m-1, n-1}.$$

Por lo tanto,

$$\delta(\mathbf{X}, \mathbf{Y}) = \mathbf{1} \{F \notin [\phi_1, \phi_2]\}, \tag{57}$$

donde ϕ_1 y ϕ_2 son tales que $\mathbb{P}(F < \phi_1) = \mathbb{P}(F > \phi_2) = \alpha/2$, es un test de nivel α para decidir entre las hipótesis $H_0 : R = 1$ contra $H_1 : R \neq 1$.

Ejemplo 7.1. Queremos construir un test de nivel $\alpha = 0.05$ para decidir entre $H_0 : \mathbb{R} = 1$ contra $H_1 : R \neq 1$ usando muestras \mathbf{X} y \mathbf{Y} de volumen $m = n = 10$.

Proponemos un test de la forma (57). El problema se reduce determinar valores ϕ_1 y ϕ_2 tales que

$$\mathbb{P}(F_{9,9} > \phi_2) = 0.025 \quad \text{y} \quad \mathbb{P}(F_{9,9} < \phi_1) = 0.025.$$

Usando las tablas de las distribuciones F resulta que $\phi_2 = 4.5362$ y que $\phi_1 = 1/\phi_2 = 0.2204$.

Finalmente, se obtiene el test

$$\delta(\mathbf{X}, \mathbf{Y}) = \{F \notin [0.2204, 4.5362]\}.$$

□

7.3. Planteo general

Supongamos que tenemos dos muestras aleatorias independientes $\mathbf{X} = (X_1, \dots, X_m)$ e $\mathbf{Y} = (Y_1, \dots, Y_n)$ con distribuciones dependientes de los parámetros ξ y η , respectivamente. Sea $\Delta = \xi - \eta$.

Se quiere decidir entre la hipótesis fundamental

$$H_0 : \Delta = \delta_0$$

contra cualquiera de las hipótesis alternativas:

- (a) $H_1 : \Delta > \delta_0;$
- (b) $H_1 : \Delta < \delta_0;$
- (c) $H_1 : \Delta \neq \delta_0.$

Sabemos que si dos estimadores para ξ y η , $\hat{\xi}_m$ y $\hat{\eta}_n$, tienen la propiedad de normalidad asintótica

$$\begin{aligned} \sqrt{m}(\hat{\xi}_m - \xi) &\rightarrow \mathcal{N}(0, \sigma^2) && \text{cuando } m \rightarrow \infty, \\ \sqrt{n}(\hat{\eta}_n - \eta) &\rightarrow \mathcal{N}(0, \tau^2) && \text{cuando } n \rightarrow \infty, \end{aligned}$$

donde σ^2 y τ^2 pueden depender de ξ y η , respectivamente y ninguna de las variables está sobre-representada (i.e., m y n son del mismo orden de magnitud), entonces

$$\frac{(\hat{\xi}_m - \hat{\eta}_n) - (\xi - \eta)}{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}} \rightarrow \mathcal{N}(0, 1) \tag{58}$$

Si σ^2 y τ^2 son conocidas, de (58) resulta que las regiones de rechazo:

$$\begin{aligned}
 \text{(a)} \quad & \frac{(\hat{\xi}_m - \hat{\eta}_n) - \delta_0}{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}} > z_{1-\alpha}; \\
 \text{(b)} \quad & \frac{(\hat{\xi}_m - \hat{\eta}_n) - \delta_0}{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}} < z_\alpha; \\
 \text{(c)} \quad & \left| \frac{(\hat{\xi}_m - \hat{\eta}_n) - \delta_0}{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}} \right| > z_{1-\alpha/2}
 \end{aligned}$$

producen un test para H_0 contra H_1 de nivel asintótico α , para cada uno de los casos considerados, respectivamente.

Si σ^2 y τ^2 son desconocidas y $\hat{\sigma}^2$ y $\hat{\tau}^2$ son estimadores consistentes para σ^2 y τ^2 , se puede demostrar que las regiones de rechazo conservan su validez cuando σ^2 y τ^2 se reemplazan por $\hat{\sigma}^2$ y $\hat{\tau}^2$, respectivamente y entonces el test con región de rechazo

$$\begin{aligned}
 \text{(a)} \quad & \frac{(\hat{\xi}_m - \hat{\eta}_n) - \delta_0}{\sqrt{\frac{\hat{\sigma}^2}{m} + \frac{\hat{\tau}^2}{n}}} > z_{1-\alpha}; \\
 \text{(b)} \quad & \frac{(\hat{\xi}_m - \hat{\eta}_n) - \delta_0}{\sqrt{\frac{\hat{\sigma}^2}{m} + \frac{\hat{\tau}^2}{n}}} < z_\alpha; \\
 \text{(c)} \quad & \left| \frac{(\hat{\xi}_m - \hat{\eta}_n) - \delta_0}{\sqrt{\frac{\hat{\sigma}^2}{m} + \frac{\hat{\tau}^2}{n}}} \right| > z_{1-\alpha/2}
 \end{aligned}$$

también tiene nivel asintótico α .

Para mayores detalles se puede consultar el libro Lehmann, E. L. (1999) *Elements of Large-Sample Theory*. Springer, New York.

Nota Bene. Notar que el argumento anterior proporciona un método general de naturaleza asintótica. En otras palabras, en la práctica los resultados que se obtienen son aproximados. Dependiendo de los casos particulares existen diversos refinamientos que permiten mejorar esta primera aproximación.

7.4. Problema de dos muestras binomiales

Sean $\mathbf{X} = (X_1, \dots, X_m)$ e $\mathbf{Y} = (Y_1, \dots, Y_n)$ dos muestras aleatorias independientes de dos variables aleatorias X e Y con distribución Bernoulli de parámetros p_X y p_Y , respectivamente. Sea $\Delta = p_X - p_Y$. Queremos un test para decidir entre las hipótesis

$$H_0 : \Delta = 0 \quad \text{contra} \quad H_1 : \Delta > 0$$

Para construir el test usaremos los estimadores de máxima verosimilitud para las probabilidades p_x y p_Y , $\hat{p}_X = \bar{X}$ y $\hat{p}_Y = \bar{Y}$.

Vamos a suponer que los volúmenes de las muestras, m y n , son suficientemente grandes y que ninguna de las dos variables está sobre representada.

Puesto que \bar{X} y \bar{Y} son estimadores consistentes para las probabilidades p_X y p_Y , resulta que los estimadores $\bar{X}(1 - \bar{X})$ y $\bar{Y}(1 - \bar{Y})$ son consistentes de las varianzas $p_X(1 - p_X)$ y $p_Y(1 - p_Y)$, respectivamente. Por lo tanto,

$$\delta(\mathbf{X}, \mathbf{Y}) = \mathbf{1} \left\{ \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{m}\bar{X}(1 - \bar{X}) + \frac{1}{n}\bar{Y}(1 - \bar{Y})}} > z_{1-\alpha} \right\}$$

es un test, de nivel aproximado α , para decidir entre las hipótesis $H_0 : \Delta = 0$ contra $H_1 : \Delta > 0$. \square

Nota Bene. Observar que el nivel del test se calcula bajo la hipótesis $p_X = p_Y$, en tal caso la desviación estándar de la diferencia $\bar{X} - \bar{Y}$ es de la forma

$$\sqrt{\frac{p_X(1 - p_X)}{m} + \frac{p_Y(1 - p_Y)}{n}} = \sqrt{p_X(1 - p_X)} \sqrt{\frac{1}{m} + \frac{1}{n}}$$

y podemos estimarla mediante

$$\sqrt{\frac{m\bar{X} + n\bar{Y}}{m+n} \left(1 - \frac{m\bar{X} + n\bar{Y}}{m+n}\right)} \sqrt{\frac{1}{m} + \frac{1}{n}}.$$

Lo que produce el test

$$\delta(\mathbf{X}, \mathbf{Y}) = \mathbf{1} \left\{ \frac{(\bar{X} - \bar{Y}) \sqrt{mn}}{\sqrt{(m\bar{X} + n\bar{Y}) \left(1 - \frac{m\bar{X} + n\bar{Y}}{m+n}\right)}} > z_{1-\alpha} \right\} \quad (59)$$

\square

Ejemplo 7.2. Se toma una muestra aleatoria de 180 argentinos y resulta que 30 están desocupados. Se toma otra muestra aleatoria de 200 uruguayos y resulta que 25 están desocupados. ¿Hay evidencia suficiente para afirmar que la tasa de desocupación de la población Argentina es superior a la del Uruguay?

Solución. La población desocupada de la Argentina puede modelarse con una variable aleatoria $X \sim \text{Bernoulli}(p_X)$ y la del Uruguay con una variable aleatoria $Y \sim \text{Bernoulli}(p_Y)$.

Para resolver el problema utilizaremos un test de nivel de significación $\alpha = 0.05$ para decidir entre las hipótesis

$$H_0 : p_X = p_Y \quad \text{contra} \quad H_1 : p_X > p_Y$$

basada en dos muestras aleatorias independientes \mathbf{X} e \mathbf{Y} de volúmenes $m = 180$ y $n = 200$, respectivamente.

El test de hipótesis dado en (59) adopta la forma

$$\delta(\mathbf{X}, \mathbf{Y}) = \mathbf{1} \left\{ \frac{(\bar{X} - \bar{Y}) \sqrt{36000}}{\sqrt{(180\bar{X} + 200\bar{Y}) \left(1 - \frac{180\bar{X}+200\bar{Y}}{380}\right)}} > 1.64 \right\} \quad (60)$$

De acuerdo con los datos observados $\bar{X} = 30/180$ y $\bar{Y} = 25/200$:

$$\frac{\left(\frac{30}{180} - \frac{25}{200}\right) \sqrt{36000}}{\sqrt{55 \left(1 - \frac{55}{380}\right)}} = 1.152\dots$$

Debido a que $1.152\dots < 1.64$, no hay evidencia suficiente para rechazar la hipótesis $p_X = p_Y$. Por lo tanto, con un 5% de nivel de significación, no hay evidencia suficiente para afirmar que la tasa de desocupación en la Argentina sea superior a la del Uruguay. \square

8. Test de la χ^2 para bondad de ajuste

8.1. Planteo del problema

Los test de bondad de ajuste tienen por objeto decidir si los datos observados se ajustan a una determinada distribución de probabilidades. Más precisamente, se formula una hipótesis, H , que afirma que los datos observados constituyen una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ de una distribución F . La distribución F puede estar completamente especificada (hipótesis simple) o puede pertenecer a una familia paramétrica (hipótesis compuesta).

Algunos ejemplos (para fijar ideas):

Ejemplo 8.1 (Moneda honesta). En una sucesión de 100 lanzamientos independientes de una moneda se observaron 55 caras y 45 cecas ¿Estos datos son compatibles con la hipótesis de que la moneda es honesta?

Ejemplo 8.2 (Multinomial). Para identificar las obras de su serie titulada *Los paisajes binarios* el artista digital Nelo firma con una imagen aleatoria de 10×10 pixels: por cada pixel lanza un dado equilibrado: si sale 1, 2 o 3 lo pinta de rojo; si sale 4 o 5 lo pinta de verde y si sale 6 lo pinta de azul. Se somete a examen la firma de una obra digital titulada *Cordillera binaria* y se obtienen los siguientes resultados: 46 pixels rojos, 37 verdes y 17 azules. ¿La obra *Cordillera binaria* pertenece a la serie *Los paisajes binarios*?

Ejemplo 8.3 (Números aleatorios). Se producen 10000 números con un generador de “números aleatorios”. Para economizar espacio se registra la cantidad de números de la forma $0.d\dots$, donde $d = 0, 1, \dots, 9$. Se obtuvieron los resultados siguientes:

d	0	1	2	3	4	5	6	7	8	9
$\#\{0.d\dots\}$	1008	1043	1014	1027	952	976	973	1021	998	988

(61)

¿Los datos se ajustan a una distribución uniforme $\mathcal{U}[0, 1]$?

Ejemplo 8.4 (Poisson). Una partícula de polen suspendida en agua es bombardeada por moléculas en movimiento térmico. Se la observa durante una hora y se registra la cantidad de impactos que recibe por segundo. Sea X la variable aleatoria que cuenta la cantidad de impactos por segundo recibidos por la partícula. Se obtuvieron los siguientes datos

X	0	1	2	3	4	5	6
# de s. con X impactos	1364	1296	642	225	55	15	3

(62)

Se quiere decidir si los datos provienen de una distribución de Poisson.

Ejemplo 8.5 (Velocidad de la luz). En la siguiente tabla se muestran las mediciones de la velocidad de la luz realizadas por el físico Albert Michelson entre el 5 de junio y el 5 de julio de 1879. Los valores dados + 299.000 son las mediciones de Michelson en km/s.

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

(63)

Las mediciones de la velocidad de la luz de Michelson, ¿se ajustan a una distribución normal?

8.2. Test de bondad de ajuste para hipótesis simples

La hipótesis nula afirma que

$$H_0 : F_X = F,$$

donde F es una distribución de probabilidades completamente determinada.

Si la hipótesis H_0 es verdadera, la función de distribución empírica, F_n de los n valores observados debe ser parecida a la función de distribución F . Lo que sugiere introducir

alguna medida de la discrepancia entre ambas distribuciones y basar el test de hipótesis en las propiedades de la distribución de dicha medida.

Hay varias formas de construir esas medidas. La que sigue fue introducida por Karl Pearson.

Se divide el rango de la variable aleatoria X en una cantidad finita k de partes disjuntas dos a dos, C_1, \dots, C_k , llamadas *clases*⁵ tales que las probabilidades $p_i = \mathbb{P}(X \in C_i | H_0) > 0$. Las k clases, C_i , serán los k conjuntos en los que agruparemos los datos para tabularlos. Se consideran n_1, \dots, n_k las frecuencias de aparición de las clases C_1, \dots, C_n en la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$,

$$n_i = \sum_{j=1}^n \mathbf{1}\{X_j \in C_i\} \quad \text{y} \quad \sum_{i=1}^k n_i = n.$$

Bajo la distribución hipotética la cantidad de valores muestrales n_i pertenecientes a la clase C_i se distribuye como una Binomial(n, p_i), y en consecuencia, para valores grandes de n , las frecuencias relativas $\frac{n_i}{n}$ deben tener valores muy próximos a las probabilidades p_i . La dispersión entre las frecuencias relativas $\frac{n_i}{n}$ y las probabilidades p_i se puede medir del siguiente modo

$$D^2 = \sum_{i=1}^k w_i \left(\frac{n_i}{n} - p_i \right)^2 = \sum_{i=1}^k w_i \frac{(n_i - np_i)^2}{n^2}, \quad (64)$$

donde los coeficientes $w_i > 0$ se pueden elegir de manera más o menos arbitraria. Cuando la hipótesis H_0 es verdadera los valores de la medida de dispersión D^2 deben ser pequeños, lo que sugiere diseñar un test de hipótesis que decida *rechazar la hipótesis H_0 cuando y solo cuando se observa que $D^2 > M$* , donde M es una constante arbitraria pero fija.

Karl Pearson demostró que cuando n es grande y la hipótesis H_0 es verdadera, poniendo $w_i = \frac{n}{p_i}$ en (64), la distribución de la medida de dispersión

$$D^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}, \quad (65)$$

es aproximadamente igual a una chi cuadrado con $k - 1$ grados de libertad. (Una demostración de este resultado puede consultarse en: Cramer, H.: Métodos matemáticos de estadística. Aguilar, Madrid. (1970).)

Test de bondad de ajuste χ^2 . Para decidir si la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ proviene de la distribución F se puede adoptar el siguiente criterio:

$$\delta(\mathbf{X}) = \mathbf{1}\{D^2 > \chi^2_{k-1, 1-\alpha}\}, \quad (66)$$

donde $\alpha \in (0, 1)$. Dicho en palabras, *rechazar que $F_X = F$ cuando y solo cuando la medida de dispersión D^2 definida en (65) supera al cuantil $1 - \alpha$ de la distribución chi cuadrado con $k - 1$ grados de libertad*. En tal caso, la probabilidad de rechazar H_0 cuando H_0 es verdadera es aproximadamente α . \square

⁵Los valores de la variable aleatoria X pertenecen a una y solo a una de las clases C_1, \dots, C_k .

8.3. Ejemplos (1^a parte)

El siguiente ejemplo tiene la virtud de mostrar, en un caso particular, una línea de demostración del resultado de Pearson sobre la distribución asintótica de D^2 .

Ejemplo 8.6 (Bernoulli). Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una distribución Bernoulli con probabilidad de éxito p . Queremos testear la hipótesis $H_0 : p = p_0$ contra $H_1 : p \neq p_0$, donde $p_0 \in (0, 1)$ es un valor determinado.

La medida de dispersión definida en (65) entre las frecuencias observadas

$$n_1 = \sum_{i=1}^n X_i \quad \text{y} \quad n_2 = n - n_1$$

y las frecuencias esperadas

$$np_0 \quad \text{y} \quad n(1 - p_0)$$

tiene la siguiente expresión

$$D^2 = \frac{(n_1 - np_0)^2}{np_0} + \frac{(n - n_1 - n(1 - p_0))^2}{n(1 - p_0)}.$$

Observando que

$$\begin{aligned} \frac{(n_1 - np_0)^2}{np_0} + \frac{(n - n_1 - n(1 - p_0))^2}{n(1 - p_0)} &= \frac{(n_1 - np_0)^2}{np_0} + \frac{(np_0 - n_1)^2}{n(1 - p_0)} \\ &= \frac{(1 - p_0)(n_1 - np_0)^2 + p_0(n_1 - np_0)^2}{np_0(1 - p_0)} \\ &= \frac{(n_1 - np_0)^2}{np_0(1 - p_0)}, \end{aligned}$$

se obtiene que

$$D^2 = \left(\frac{n_1 - np_0}{\sqrt{np_0(1 - p_0)}} \right)^2 \tag{67}$$

Cuando la hipótesis H_0 es verdadera, $n_1 \sim \text{Binomial}(n, p_0)$, y de acuerdo con el teorema central del límite la distribución de la variable aleatoria

$$\frac{n_1 - np_0}{\sqrt{np_0(1 - p_0)}}$$

es asintóticamente normal $\mathcal{N}(0, 1)$. Por lo tanto, para valores grandes de n , D^2 tiene una distribución aproximadamente igual a χ_1^2 . \square

Ejemplo 8.1. (Continuación) Se trata de un caso particular del esquema anterior, donde $p_0 = 1/2$ y $n = 100$. En consecuencia, la medida de dispersión (67) es

$$D^2 = \left(\frac{n_1 - 50}{5} \right)^2,$$

y para un nivel de significación α el test de hipótesis (66) adopta la forma

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ \left(\frac{n_1 - 50}{5} \right)^2 > \chi_{1, 1-\alpha}^2 \right\}.$$

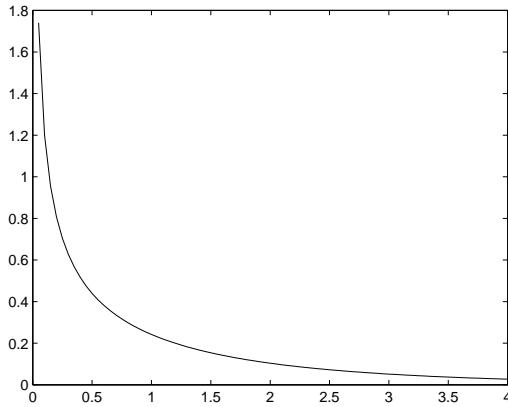


Figura 11: La densidad χ_1^2 .

Consultado la tabla de cuantiles de la distribución χ_1^2 vemos que $\chi_{1, 0.95}^2 = 3.841$.

De acuerdo con los datos observados $n_1 = 55$, de donde sigue que como $D^2 = \left(\frac{55-50}{5} \right)^2 = 1$. En vista de que $1 < \chi_{1, 0.95}^2$, a un nivel de significación del 5% el test no rechaza la hipótesis de que se la moneda sea honesta. \square

Ejemplo 8.2. (Continuación) El color en cada pixel se modela con una variable aleatoria X a valores $\{r, g, b\}$ cuya distribución está completamente determinada por los valores de las probabilidades $\mathbb{P}(X = r) = p_r$, $\mathbb{P}(X = g) = p_g$ y $\mathbb{P}(X = b) = p_b$. Queremos decidir si los datos obtenidos son compatibles (o no) con la hipótesis

$$H_0 : p_r = 3/6, p_g = 2/6, p_b = 1/6.$$

Para ello construimos un test de bondad de ajuste basado en una muestra aleatoria, $\mathbf{X} = (X_1, \dots, X_n)$ de volumen $n = 10 \times 10 = 100$. Prescrito el nivel de significación α y clasificando los datos de acuerdo con el color observado obtenemos un test de la forma

$$\delta(\mathbf{X}) = \mathbf{1}\{D^2 > \chi_{2, 1-\alpha}^2\},$$

donde

$$D^2 = \frac{(n_r - 100(3/6))^2}{100(3/6)} + \frac{(n_g - 100(2/6))^2}{100(2/6)} + \frac{(n_b - 100(1/6))^2}{100(1/6)}.$$

Por ejemplo, si se prescribe un nivel de significación del 1% (i.e., $\alpha = 0.01$) tenemos que $\chi_{2,1-\alpha}^2 = \chi_{2,0.99}^2 = 9.2103$ y el test adopta la forma

$$\delta(\mathbf{X}) = \mathbf{1} \left\{ \frac{(n_r - 50)^2}{50} + \frac{(n_g - 33.33...)^2}{33.33...} + \frac{(n_b - 16.66...)^2}{16.66...} > 9.2103 \right\},$$

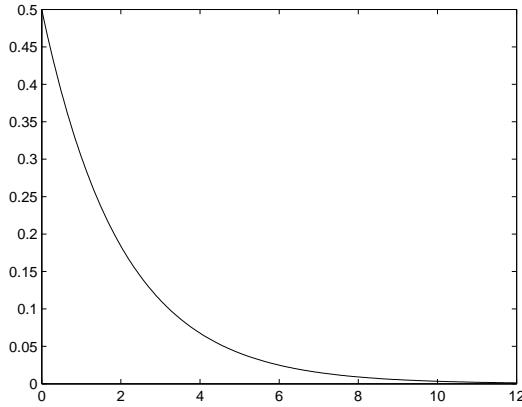


Figura 12: La densidad χ_2^2 .

De acuerdo con los datos observados: $n_r = 46$, $n_g = 37$ y $n_b = 17$ y la medida de dispersión de Pearson vale

$$D^2 = \frac{(46 - 50)^2}{50} + \frac{(37 - 33.33...)^2}{33.33...} + \frac{(17 - 16.66...)^2}{16.66...} = 0.73$$

Motivo por el cual, no hay evidencia que permita rechazar que la obra *Cordillera binaria* pertenece a la serie *Los paisajes binarios* del artista Nelo.

Notar que para rechazar que la obra citada pertenece al artista se necesitaba un test de la forma $\delta(\mathbf{X}) = \{D^2 \geq 0.73\}$. Bajo la hipótesis H_0 , $D^2 \sim \chi_2^2$ y $p = \mathbb{P}(D^2 \geq 0.73) = 0.694\dots$ y en ese caso, la probabilidad de equivocarse al rechazar que la obra pertenece a Nelo es del orden del 69%. \square

Ejemplo 8.3. (Continuación) En este caso las clases C_i son los intervalos de la forma $[\frac{i-1}{10}, \frac{i}{10})$, $i = 1, \dots, 10$. Si la variable aleatoria X tuviese distribución $\mathcal{U}[0, 1]$, $p_i = \mathbb{P}(X \in C_i) = 1/10$. El volumen de la muestra es $n = 10000$. Las frecuencias observadas, n_i , son los valores que se muestran en la tabla (61). Las frecuencias esperadas, np_i , son todas iguales y valen 1000. Por lo tanto, la medida de dispersión de Pearson vale

$$D^2 = \frac{1}{1000} (8^2 + 43^2 + 14^2 + 27^2 + 48^2 + 24^2 + 27^2 + 21^2 + 2^2 + 12^2) = 7.036$$

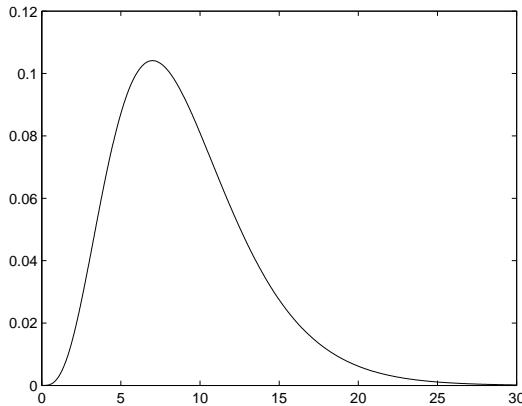


Figura 13: La densidad χ^2_9 . El área bajo la curva a la derecha del valor 7.036 es 0.6336....

Bajo la hipótesis $X \sim \mathcal{U}[0, 1]$, la medida de dispersión D^2 se distribuye como una chi cuadrado con 9 grados de libertad. Si se observa la Figura 13 se puede ver que un valor de 7.036 para D^2 no es inusual, lo que indica que no hay evidencia suficiente para rechazar la hipótesis $X \sim \mathcal{U}[0, 1]$. Para rechazar dicha hipótesis se necesita un test de la forma $\delta(\mathbf{X}) = \{D^2 \geq 7.036\}$. Bajo la hipótesis $X \sim \mathcal{U}[0, 1]$, $p = \mathbb{P}(D^2 \geq 7.036) = 0.6336\dots$ y en tal caso, la probabilidad de equivocarse al rechazar que los datos provienen de una distribución uniforme es del orden del 63 %. \square

8.4. Comentarios sobre el método

En la sección 8.2 presentamos el test de bondad de ajuste χ^2 de Pearson. En la sección 8.3 ilustramos su implementación en algunos ejemplos muy simples. Esos ejemplos comparten una característica en común: las clases en que dividimos el rango de la variable X estaban condicionadas por el modo en que estaban tabulados los datos observados.

Esos ejemplos podrían oscurecer el siguiente hecho que no puede pasar desapercibido: *el procedimiento de construcción de las clases C_1, \dots, C_k en que se divide el rango de la variable es (más o menos) arbitrario*. En la descripción del método presentada en la sección 8.2 no se indica cuántas clases deben considerarse ni se indica cómo deben ser esas clases.

Sobre la cantidad de clases (1). Un lector desprevenido podría pensar que para implementar el método basta dividir el rango de la variable en dos clases. Ese modo de proceder no es recomendable. ¿Usando las clases, $C_1 = [-1, 0]$ y $C_2 = (0, 1]$, podrían distinguirse la distribución uniforme sobre el $[-1, 1]$ de la distribución triangular con el mismo soporte? Evidentemente no. Sin embargo, en cuanto aumentamos la cantidad de clases, a 4 por ejemplo, la diferencia se podría percibir.

Cuando agrupamos los datos en clases y conservamos solamente la frecuencia con que

se observa cada clase destruimos información sobre la variable muestreada. Si la cantidad de partes es muy chica, se pierde mucha información y la resolución del test es bastante mala. \square

Sobre la cantidad y la forma de las clases (2). \heartsuit Se podría pensar que al aumentar la cantidad de clases en que se divide el rango de la variable mejora la resolución del test, esto es parcialmente correcto. Si nos excedemos en la cantidad de clases la distribución de la medida de dispersión D^2 deja de parecerse a la χ^2 .

Debido a su naturaleza asintótica, el test de bondad de ajuste χ^2 funciona bien solamente cuando las frecuencias esperadas en todas las clases es relativamente grande. En la Bibliografía consultada no se comenta ningún método “óptimo” para determinar la cantidad de clases en que debe dividirse el rango de la variable aleatoria. Aunque sobre este asunto parece no existir acuerdo entre los especialistas, todos coinciden en que la cantidad de clases está limitada por una condición del siguiente tipo:

- $np_i \geq 5$ para $i = 1, \dots, k$ (Fisher);
- $np_i \geq 10$ para $i = 1, \dots, k$ (Cramer);
- $np_i \geq 8$ para $i = 1, \dots, k$ (Borovkov).

DeGroot indica que la condición de Fisher es suficiente para que la distribución χ^2 sea una *buenas aproximación* de la distribución de D^2 . Incluso afirma que, poniendo $np_i > 1.5$ la aproximación continua siendo satisfactoria. \square

En todo lo que sigue adoptaremos la condición de Cramer sobre la cantidad y forma de las clases: $np_i \geq 10$ para $i = 1, \dots, k$. De este modo, si para algún i ocurriese que $np_i < 10$ redefinimos la partición C_1, \dots, C_k del rango de la variable. Por ejemplo, uniendo C_i con C_{i+1} . Esta condición implica que si el volumen de la muestra no es muy grande, la partición del rango de la variable no puede ser muy fina.

Ejemplo 8.7 (Exponencial). Se dispone de los siguientes datos sobre la duración en horas de 100 baterías:

3.9662191	0.5819433	0.1842986	0.5977917	1.9781844
0.6048519	0.7259459	1.5896094	0.2411217	2.4502631
1.6993148	0.9884268	0.4281823	2.0079459	0.0022114
0.0422904	1.6384416	0.2214073	0.4350003	0.1934794
0.3548681	0.7775309	0.1052627	0.6497803	0.7227835
3.0542040	3.4097021	0.3577800	1.4532404	2.2825177
1.4903543	0.6062705	0.9444304	0.1119637	1.2789623
0.3598502	0.8901427	0.1282656	0.3331565	1.6096607
1.3348741	3.1158026	0.4525998	0.4554032	0.8698826
0.0215405	0.7115861	0.4859616	1.3781469	0.0979241
0.8608390	0.1999889	0.6616866	0.6960469	1.4041375
1.6087253	0.2149426	0.4833662	2.3159498	1.0346222

0.2056717	0.5228204	1.8704697	0.2166610	0.9409121
3.4983549	0.3543629	1.5233421	0.1877053	0.3911424
0.1840173	1.1453108	0.0161651	1.7702696	1.0397349
0.0772446	0.0421012	0.4814322	2.5107661	1.6500077
1.2448903	0.1030540	0.4572152	0.6299386	0.1021735
0.2197928	1.1234052	0.0936486	1.6546837	3.1267264
1.4791009	0.3132625	1.0092715	1.2217523	3.2381804
0.1215625	0.7677260	0.2124635	2.2532736	0.7156024

¿Puede afirmarse a un nivel del 1% que la duración de las baterías se ajusta a una distribución exponencial de media 2 horas?

Solución.

1. *Construyendo una partición.* Lo primero que tenemos que hacer es determinar la cantidad y la forma de las clases en que agruparemos los datos.

Con la indicación de Cramer ($np_i \geq 10$, para $i = 1, \dots, k$) la máxima cantidad de clases que podemos elegir es 10. Para simplificar un poco las cuentas elegiremos una partición en 7 clases, C_1, \dots, C_7 , que sean equiprobables bajo la distribución hipotética: $X \sim \text{Exponencial}(1/2)$.⁶

Cuando la función de distribución de una variable aleatoria es continua la construcción de la partición en k clases equiprobables se resuelve utilizando los cuantiles. La clase C_i será el intervalo $\left[x_{\frac{i-1}{k}}, x_{\frac{i}{k}}\right)$, donde $x_{\frac{i}{k}}$ es el cuantil- $\frac{i}{k}$ de la distribución hipotética.

La función de distribución de la exponencial de media 2 es $F(x) = (1 - e^{-x/2})\mathbf{1}\{x \geq 0\}$ y su cuantil- γ es la única solución de la ecuación $F(x_\gamma) = \gamma$. En consecuencia, $x_\gamma = -2 \log(1 - \gamma)$. En consecuencia, para obtener 7 clases equiprobables basta poner

$$C_i = \left[-2 \log \left(1 - \frac{i-1}{7} \right), -2 \log \left(1 - \frac{i}{7} \right) \right), \quad i = 1, \dots, 7,$$

lo que produce: $C_1 = [0, 0.3083)$, $C_2 = [0.3083, 0.6729)$, $C_3 = [0.6729, 1.1192)$, $C_4 = [1.1192, 1.6946)$, $C_5 = [1.6946, 2.5055)$, $C_6 = [2.5055, 3.8918)$ y $C_7 = [3.8918, \infty)$.

2. *Agrupando los datos.* Determinadas las clases agrupamos los datos. En la siguiente tabla se muestran las frecuencias observadas y la cantidad que aporta cada clase a la medida de dispersión D^2 :

n_i	26	23	16	18	9	7	1
$(n_i - np_i)^2 / np_i$	9.60571	5.31571	0.20571	0.96571	1.95571	3.71571	12.35571

3. *Decisión al 1%.* Finalmente comparamos el valor obtenido para $D^2 = 34.12$ con el cuantil 0.99 de la distribución $\chi^2_{6,0.99} = 16.812$. Como $D^2 > \chi^2_{6,0.99}$ concluimos que la duración de las pilas no se ajusta a la distribución exponencial de media 2 horas. \square

⁶Notar que al elegir el criterio de las clases “equiprobables” para construir la partición, garantizamos de entrada que no habrá partes sub o sobre dimensionadas y no vamos a encontrarnos con el problema de tener que unir dos clases porque quedaron muy “flacas”.

Nota Bene. No siempre se puede dividir el rango de la variable en clases de igual probabilidad. Las variables discretas no lo permiten. En tal caso habrá que conformarse con algunas partes suficientemente “gorditas” como para que valga la condición $np_i \geq 10$ \square

8.5. Test de bondad de ajuste para hipótesis compuestas

La hipótesis nula afirma que

$$H_0 : F_X = F_{\theta_1, \dots, \theta_r},$$

donde $F_{\theta_1, \dots, \theta_r}$ es una distribución de probabilidades perteneciente a una familia paramétrica completamente determinada y los valores de los parámetros $\theta_1, \dots, \theta_r$ son desconocidos.

En este caso los r parámetros desconocidos se estiman usando el método de máxima verosimilitud. Los valores de las r estimaciones se “enchufan” en la distribución paramétrica como si fuesen los verdaderos valores de los parámetros y se aplica el test χ^2 desarrollado en la sección 8.2. Solo que ahora se perderá un grado de libertad por cada parámetro estimado. Si para construir la medida de dispersión D^2 se recurrió a una partición del rango de la variable X en k clases, la distribución de D^2 será aproximadamente una χ^2_{k-1-r} .

Ejemplo 8.4. (Continuación) La hipótesis H_0 afirma que la cantidad de impactos por segundo recibidos por la partícula de polen sigue una distribución de Poisson, pero no indica cuál es su media (el parámetro λ).

El estimador de máxima verosimilitud para la media de una distribución de Poisson es $\hat{\lambda}_{mv} = \bar{X}$. Usando los datos que aparecen en la tabla (62) obtenemos

$$\hat{\lambda}_{mv} = \frac{0(1364) + 1(1296) + 2(642) + 3(225) + 4(55) + 5(15) + 6(3)}{3600} = \frac{3568}{3600} = 0.9911 \approx 1.$$

Las clases C_i se pueden construir usando como criterio que $3600\mathbb{P}(X \in C_i) \geq 10$. Si suponemos que $X \sim \text{Poisson}(1)$, su función de probabilidades será $\mathbb{P}(X = n) = e^{-1}/n!$, $n = 0, 1, \dots$

Usaremos como partición las siguientes clases: $C_1 = \{0\}$, $C_2 = \{1\}$, $C_3 = \{2\}$, $C_4 = \{3, 4, 5, \dots\}$, cuyas probabilidades son $p_1 = p_2 = 0.3678$, $p_3 = 0.1839$ y $p_4 = 0.0805$. Obtenemos que

$$\begin{aligned} D^2 &= \frac{(1364 - 3600p_1)^2}{3600p_1} + \frac{(1296 - 3600p_2)^2}{3600p_2} + \frac{(642 - 3600p_3)^2}{3600p_3} + \frac{(298 - 3600p_4)^2}{3600p_4} \\ &= \frac{1593.6064}{1324.08} + \frac{788.4864}{1324.08} + \frac{401.6016}{662.04} + \frac{67.24}{289.8} = 2.6376 \end{aligned}$$

Si se observa la Figura 12 se puede ver que un valor de 2.6376 para D^2 no es inusual para una distribución χ^2_2 , lo que indica que la cantidad de impactos recibidos por la partícula de polen se puede considerar como una variable aleatoria con distribución Poisson. \square

Ejemplo 8.5. (Continuación) La hipótesis nula es de la forma $H_0 : X \sim \mathcal{N}(\mu, \sigma^2)$. Informalmente, se puede ver usando un histograma que los datos “obedecen” a una distribución normal.

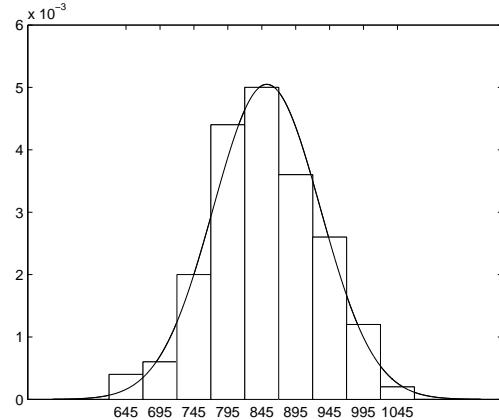


Figura 14: Histograma de los mediciones de Michelson y gráfico de la densidad de la distribución de media $\bar{X} = 852.4$ y varianza $S^2 = 79.0105$.

Usando los cuantiles de la distribución normal de media 852.4 y varianza 79.0105, construimos 9 clases equiprobables delimitadas por los valores: 756, 792, 818, 841, 863, 886, 913 y 949. Las frecuencias observadas en cada una de las 9 clases son, respectivamente, 9, 11, 15, 12, 11, 14, 7, 6 y 15. Con esos datos, la medida de dispersión resulta $D^2 = 7.82 < \chi^2_{6, 0.90} \dots$ □

9. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Bolfarine, H., Sandoval, M. C.: Introdução à Inferência Estatística. SBM, Rio de Janeiro. (2001).
2. Borovkov, A. A.: Estadística matemática. Mir, Moscú. (1984).
3. Cramer, H.: Métodos matemáticos de estadística. Aguilar, Madrid. (1970).
4. DeGroot, M. H.: Probability and Statistics. Addison-Wesley, Massachusetts. (1986).
5. Fisher, R. A.: Statistical methods for research workers. Hafner, New York (1954).
6. Hoel P. G.: Introducción a la estadística matemática. Ariel, Barcelona. (1980).
7. Lehmann, E. L.: Elements of Large-Sample Theory. Springer, New York. (1999)

8. Maronna R.: Probabilidad y Estadística Elementales para Estudiantes de Ciencias. Editorial Exacta, La Plata. (1995).
9. Meyer, P. L.: Introductory Probability and Statistical Applications. Addison-Wesley, Massachusetts. (1972).
10. Rice, J. A.: Mathematical Statistics and Data Analysis. Duxbury Press, Belmont. (1995).
11. Ross, S. M.: Introduction to Probability and Statistics for Engineers and Scientists. Elsevier Academic Press, San Diego. (2004)
12. Walpole, R. E.: Probabilidad y estadística para ingenieros, 6a. ed., Prentice Hall, México. (1998)

Análisis Bayesiano (Borradores, Curso 23)

Sebastian Grynberg

17-19 de junio de 2013



*Aquí no valen Dotores,
Solo vale la experiencia,
Aquí verían su inocencia
Esos que todo lo saben;
Por que esto tiene otra llave
Y el gaucho tiene su ciencia.*
(Martín Fierro)

Índice

1. Análisis Bayesiano	2
1.1. Distribuciones <i>a priori</i> y <i>a posteriori</i>	2
1.2. Distribuciones predictivas	5
1.3. Estimadores Bayesianos	6
1.4. Estimación por intervalo para parámetro continuo	6
1.5. Sobre la distribución a priori uniforme.	7
2. Ejemplos	8
2.1. Las distribuciones β y el problema del “control de calidad”	8
2.2. Normales de varianza conocida y media normal	13
2.3. Distribuciones Poisson con a priori Gamma	16
3. Bibliografía consultada	19

1. Análisis Bayesiano

Si se lo compara con el modelado probabilístico, el propósito del análisis estadístico es fundamentalmente un propósito de *inversión*, ya que se propone inferir las causas (los parámetros del mecanismo aleatorio) a partir de los efectos (las observaciones). En otras palabras, cuando observamos un fenómeno aleatorio regulado por un parámetro θ , los métodos estadísticos nos permiten deducir de las observaciones una *inferencia* (esto es, un resumen, una caracterización) sobre θ , mientras que el modelado probabilístico caracteriza el comportamiento de las observaciones futuras *condicionales* a θ . Este aspecto de la estadística es obvio en la noción de función de verosimilitud, puesto que, formalmente, es la densidad conjunta de la muestra reescrita en el orden propio

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta), \quad (1)$$

i.e., como una función de θ , que es *desconocida*, que depende de los valores observados \mathbf{x} .

La *regla de Bayes* es una descripción general de la inversión de probabilidades: si A y E son eventos de probabilidad positiva, $\mathbb{P}(A|E)$ y $\mathbb{P}(E|A)$ están relacionados por

$$\mathbb{P}(A|E) = \frac{\mathbb{P}(E|A)\mathbb{P}(A)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|A)\mathbb{P}(A)}{\mathbb{P}(E|A)\mathbb{P}(A) + \mathbb{P}(E|A^c)\mathbb{P}(A^c)}.$$

En su versión continua, la regla de Bayes establece que dadas dos variables aleatorias X e Y , con distribución condicional $f_{X|Y=y}(x)$ y distribución marginal $f_Y(y)$, la distribución condicional de Y dado que $X = x$ es

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)f_Y(y)}{\int f_{X|Y=y}(x)f_Y(y)dy}.$$

1.1. Distribuciones *a priori* y *a posteriori*

Desde el punto de vista probabilístico el teorema de inversión es bastante natural. Bayes y Laplace fueron más allá y consideraron que la incertezas sobre el parámetro desconocido de

un modelo paramétrico puede modelarse mediante una distribución de probabilidad sobre el espacio paramétrico.

La esencia del enfoque Bayesiano consiste en que el parámetro desconocido, θ , se considera como *variable aleatoria* con cierta función densidad de probabilidades

$$\pi_\theta(t), \quad t \in \Theta.$$

La densidad $\pi_\theta(t)$ se llama densidad *a priori*, o sea, dada *antes* del experimento. El enfoque Bayesiano supone que el parámetro desconocido θ se ha escogido aleatoriamente de la distribución cuya densidad es $\pi_\theta(t)$.

Definición 1.1. Un modelo estadístico Bayesiano está hecho de un modelo paramétrico $\mathcal{F} = \{f(x|t) : t \in \Theta\}$ para las observaciones y una distribución de probabilidad *a priori* $\pi_\theta(t)$ sobre el espacio paramétrico Θ .

Nota Bene. En un modelo Bayesiano, la “densidad” muestral $f(x|t)$, $t \in \Theta$, es la “densidad” condicional de la variable aleatoria X dado que $\theta = t$.

Dado un modelo Bayesiano podemos construir varias distribuciones, a saber:

1. La distribución *conjunta* del parámetro θ y la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$:

$$f_{\theta, \mathbf{X}}(t, \mathbf{x}) = f(\mathbf{x}|t)\pi_\theta(t) = \left(\prod_{i=1}^n f(x_i|t) \right) \pi_\theta(t). \quad (2)$$

2. La distribución *marginal* de la muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$:

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{\Theta} f_{\theta, \mathbf{X}}(t, \mathbf{x}) dt = \int_{\Theta} f(\mathbf{x}|t)\pi_\theta(t) dt. \quad (3)$$

3. La distribución *a posteriori* (o sea, *después del experimento*) de la variable aleatoria θ , obtenida mediante la fórmula de Bayes:

$$\pi(t|\mathbf{x}) = \frac{f_{\theta, \mathbf{X}}(t, \mathbf{x})}{\int_{\Theta} f_{\theta, \mathbf{X}}(t, \mathbf{x}) dt} = \frac{f(\mathbf{x}|t)\pi_\theta(t)}{\int_{\Theta} f(\mathbf{x}|t)\pi_\theta(t) dt}. \quad (4)$$

Nota Bene. Si el parámetro θ es una variable aleatoria discreta, la “densidad” a priori $\pi_\theta(t)$ debe interpretarse como la función de probabilidades y las expresiones del tipo $\int dt$ deben reemplazarse por expresiones del tipo \sum_t .

Ejemplo 1.2 (Bayes (1764)). Se echa a rodar una bola de billar B_1 sobre una línea de longitud 1, con probabilidad uniforme de que se detenga en cualquier lugar. Se detiene en θ . Una segunda bola B_2 se echa a rodar 5 veces bajo las mismas condiciones que la primera y X denota la cantidad de veces que la bola B_2 se detuvo a la izquierda de donde lo hizo B_1 . Dado que $X = x$, ¿qué se puede inferir sobre θ ?

El problema consiste en hallar la distribución *a posteriori* de θ dado que $X = x$, cuando la distribución *a priori* de θ es uniforme sobre $(0, 1)$ y $X \sim \text{Binomial}(5, \theta)$. Puesto que

$$f(x|t) = \binom{5}{x} t^x (1-t)^{5-x} \quad \text{y} \quad \pi_\theta(t) = \mathbf{1}\{t \in (0, 1)\},$$

la distribución conjunta del parámetro θ y la variable aleatoria X es

$$f_{\theta,X}(t, x) = \binom{5}{x} t^x (1-t)^{5-x} \mathbf{1}\{t \in (0, 1)\}$$

y la distribución marginal de la variable X es

$$\begin{aligned} f_X(x) &= \int_0^1 \binom{5}{x} t^x (1-t)^{5-x} dt = \binom{5}{x} \int_0^1 t^x (1-t)^{5-x} dt = \binom{5}{x} \frac{\Gamma(x+1)\Gamma(6-x)}{\Gamma(7)} \\ &= \frac{5!}{x!(5-x)!} \frac{x!(5-x)!}{6!} = \frac{1}{6}, \quad x = 0, 1, \dots, 5 \end{aligned}$$

(En palabras, los 6 posibles valores de X son igualmente probables.)

De lo anterior se deduce que la distribución a posteriori de θ dado que $X = x$

$$\pi(t|x) = 6 \binom{5}{x} t^x (1-t)^{5-x} \mathbf{1}\{t \in (0, 1)\},$$

i.e., la distribución de θ condicional a que $X = x$ es la distribución $\beta(x+1, 6-x)$. \square

Ejemplo 1.3 (Laplace (1773)). En una urna hay 12 bolas blancas y negras. Si la primer bola extraída es blanca, ¿cuál es la probabilidad de que la proporción θ de bolas blancas sea $2/3$? Asumiendo *a priori* que las cantidades 2 a 11 de bolas blancas son igualmente probables, i.e., que θ es equiprobable sobre $\{2/12, \dots, 11/12\}$. La distribución a posteriori de θ se deduce usando el teorema de Bayes:

$$\pi(2/3|\text{datos}) = \frac{(2/3)(1/10)}{\sum_{p=2/12}^{11/12} p(1/10)} = \frac{(2/3)}{\sum_{n=2}^{11} n/12} = \frac{8}{(11 \times 12)/2 - 1} = \frac{8}{65}.$$

\square

Principio de verosimilitud. La fórmula de Bayes (4) puede leerse del siguiente modo: observado que la muestra aleatoria \mathbf{X} arrojó los valores \mathbf{x} , la distribución a posteriori de θ es proporcional a la función de verosimilitud $L(t|\mathbf{x}) = f(\mathbf{x}|t)$ multiplicada por la distribución a priori de θ . En símbolos

$$\pi(t|\mathbf{x}) \propto L(t|\mathbf{x})\pi_\theta(t).$$

Esto significa que la información sobre la variable θ que viene en una muestra \mathbf{x} está completamente contenida en la función de verosimilitud $L(t|\mathbf{x})$. Más aún, cuando \mathbf{x}_1 y \mathbf{x}_2 son dos observaciones que dependen del mismo parámetro θ y existe una constante c que satisface

$$L_1(t|\mathbf{x}_1) = cL_2(t|\mathbf{x}_2)$$

para cada $t \in \Theta$, entonces \mathbf{x}_1 y \mathbf{x}_2 tienen la misma información sobre θ y deben conducir a inferencias idénticas. Esto es así porque el análisis Bayesiano se basa completamente en la distribución *a posteriori* $\pi(t|\mathbf{x})$ que depende de \mathbf{x} solo a través de $L(t|\mathbf{x})$. \square

Ejemplo 1.4. Trabajando sobre el ranking de una serie televisiva un investigador encontró 9 espectadores que la miran y 3 que no la miran. Si no se dispone de más información sobre el experimento, se pueden proponer al menos dos modelos. Si $\theta \in (0, 1)$ representa la proporción de los espectadores que mira la serie:

(1) El investigador encuestó a 12 personas y por lo tanto observó $X \sim \text{Binomial}(12, \theta)$ con $X = 9$.

(2) El investigador encuestó Y personas hasta que encontró 3 que no miraban la serie y por lo tanto observó $Y \sim \text{Pascal}(3, 1 - \theta)$ con $Y = 12$.

El punto importante es que, en cualquiera de los dos modelos, la verosimilitud es proporcional a

$$\theta^3(1 - \theta)^9.$$

Por lo tanto, el principio de verosimilitud implica que la inferencia sobre θ debe ser idéntica para ambos modelos. \square

1.2. Distribuciones predictivas

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una distribución indexada por θ . Se observa que $\mathbf{X} = \mathbf{x}$ y se quiere *predecir* una el comportamiento de una nueva observación $Y \sim g(y|\theta)$, donde Y es una variable aleatoria que depende del mismo parámetro θ . En el contexto probabilístico *predecir* significa contestar preguntas del tipo: ¿con qué probabilidad se observaran valores en un intervalo dado? En otras palabras ¿cuál será la distribución de la nueva observación Y ?

Este problema se puede resolver usando la fórmula de probabilidad total. Dado que se observó $\mathbf{X} = \mathbf{x}$, la función *densidad predictiva* (*o incondicional*) de la nueva observación Y será

$$g(y|\mathbf{x}) = \int g(y|t)\pi(t|\mathbf{x})dt. \quad (5)$$

El primer factor del integrando que aparece en (5) corresponde a las densidades de la variable aleatoria Y condicionadas al conocimiento de que $\theta = t$. El segundo factor corresponde a la densidad a posteriori del parámetro aleatorio θ .

Si tuviésemos la capacidad de observar qué valor arrojó la variable θ y observáramos que $\theta = t$, la predicción de Y quedaría determinada por la *densidad condicional* $g(y|t)$. Sin embargo, la hipótesis fundamental de este enfoque es que el parámetro θ no puede ser observado y lo único que podemos observar es la muestra aleatoria \mathbf{X} . El calificativo de *incondicional* que se le otorga a la densidad $g(y|\mathbf{x})$ obtenida en (5) está puesto para destacar que su construcción no utiliza observaciones del parámetro θ .

Ejemplo 1.5 (Bayes (1764) Continuación.). Supongamos ahora que la bola B_2 se detuvo exactamente 3 veces a la izquierda de donde lo hizo la bola B_1 , ¿cuál es la probabilidad p de que al echar a rodar una tercera bola de billar B_3 también se detenga a la izquierda de donde se detuvo B_1 ?

Sea $Y \sim \text{Bernoulli}(\theta)$ la variable aleatoria que vale 1 si la bola B_3 se detiene a la izquierda de donde se detuvo B_1 y 0 en caso contrario. Para calcular p usamos la distribución predictiva:

$$p = \mathbb{P}(Y = 1|X = 3) = \int_0^1 \mathbb{P}(Y = 1|t)\pi(t|3)dt = \int_0^1 t\pi(t|3) = \mathbb{E}[\theta|X = 3].$$

Como $\theta|X = 3 \sim \beta(4, 2)$, resulta que $p = 4/6$. \square

1.3. Estimadores Bayesianos

1. **Estimación bayesiana por esperanza condicional.** En el contexto Bayesiano θ es una variable aleatoria. Entre todas las funciones (de la muestra aleatoria \mathbf{X}) $\hat{\theta} = \varphi(\mathbf{X})$ la mejor estimación para θ (desde el punto de vista de minimizar el error cuadrático medio $\mathbb{E}[(\theta - \varphi(\mathbf{X}))^2]$) es la esperanza condicional $\mathbb{E}[\theta|\mathbf{X}]$:

$$\hat{\theta}(\mathbf{X}) = \mathbb{E}[\theta|\mathbf{X}] = \int t\pi(t|\mathbf{X})dt. \quad (6)$$

2. **Estimación bayesiana por máximo a posteriori.** Otro estimador, de uso frecuente, es el llamado *máximo a posteriori* (*o moda*) definido por

$$\hat{\theta}_{map}(\mathbf{X}) := \arg \max_{t \in \Theta} \pi(t|\mathbf{X}). \quad (7)$$

Ejemplo 1.6 (Bayes (1764) Continuación.). Supongamos ahora que la bola B_2 se detuvo exactamente 3 veces a la izquierda de donde lo hizo la bola B_1 . En tal caso

$$\hat{\theta}(3) = \mathbb{E}[\theta|X = 3] = \frac{4}{6}$$

y

$$\hat{\theta}_{map}(3) = \arg \max_{t \in (0,1)} 6 \binom{5}{3} t^3 (1-t)^2 = \arg \max_{t \in (0,1)} t^3 (1-t)^2.$$

Como el logaritmo es una función creciente, el argumento que maximiza a la función $t^3(1-t)^2$ coincide con el argumento maximizador de la función $\psi(t) = \log(t^3(1-t)^2) = 3\log(t) + 2\log(1-t)$. Observando que

$$0 = \frac{d}{dt} \psi(t) = \frac{3}{t} - \frac{2}{1-t} \iff 3(1-t) - 2t = 0 \iff t = \frac{3}{5},$$

se puede deducir que

$$\hat{\theta}_{map}(3) = \frac{3}{5}.$$

□

1.4. Estimación por intervalo para parámetro continuo

Dada la muestra aleatoria \mathbf{X} se desea construir intervalos (acotados) que capturen casi toda la variabilidad del parámetro aleatorio θ . Si el intervalo $[a, b]$ es tal que

$$\mathbb{P}(\theta \in [a, b]|\mathbf{X}) = 1 - \alpha, \quad (8)$$

será llamado *intervalo estimador de nivel* $1 - \alpha$. En la práctica, los valores de α son pequeños: 0.1 o 0.05 o 0.01. En general, los valores de a y b dependerán de los valores de la muestra aleatoria \mathbf{x} . Dado que $\mathbf{X} = \mathbf{x}$, los intervalos estimadores de nivel $1 - \alpha$ se obtienen resolviendo la siguiente ecuación de las variables a y b :

$$\int_a^b \pi(t|\mathbf{x})dt = 1 - \alpha. \quad (9)$$

De todas las soluciones posibles de la ecuación (9) se prefieren aquellas que producen intervalos de longitud lo más pequeña posible.

Una solución particular de la ecuación (9) puede obtenerse mediante el siguiente razonamiento: como la distribución a posteriori del parámetro θ está centrada alrededor de su esperanza, $\hat{\theta}(\mathbf{x}) := \mathbb{E}[\theta|\mathbf{X} = \mathbf{x}]$, y no puede desviarse demasiado de allí, los intervalos que la contengan deben ser relativamente pequeños. Esto sugiere la siguiente construcción: dividir a la mitad el nivel y tratar de capturar cada una de las mitades a izquierda y a derecha de $\hat{\theta}(\mathbf{x})$. En otras palabras, se trata de resolver las siguientes ecuaciones:

$$\int_a^{\hat{\theta}(\mathbf{x})} \pi(t|\mathbf{x})dt = \frac{1-\alpha}{2}, \quad \int_{\hat{\theta}(\mathbf{x})}^b \pi(t|\mathbf{x})dt = \frac{1-\alpha}{2}. \quad (10)$$

Ejemplo 1.7. Se considera el siguiente modelo Bayesiano: $X \sim \mathcal{N}(\theta, 1)$ con distribución a priori $\theta \sim \mathcal{N}(0, 10)$. Sobre la base de una muestra de tamaño 1 de X se quiere determinar un intervalo de nivel $1 - \alpha$ para la variable θ .

Dado que $X = x$ tenemos que

$$\pi(t|x) \propto L(\theta|x)\pi_\theta(t) \propto \exp\left(-\frac{(x-t)^2}{2} - \frac{t^2}{20}\right) \propto \exp\left(-\frac{11}{20}\left(t - \frac{10x}{11}\right)^2\right)$$

y por lo tanto $\theta|X = x \sim \mathcal{N}\left(\frac{10x}{11}, \frac{10}{11}\right)$. Como la variable

$$Z = \frac{(\theta|X = x) - (10x/11)}{\sqrt{10/11}} \sim \mathcal{N}(0, 1)$$

tenemos que $\mathbb{P}(|Z| < z_{1-\alpha/2}) = 1 - \alpha$ y de allí se deduce dado que $X = x$ el intervalo

$$\left[\frac{10x}{11} - z_{1-\alpha/2} \sqrt{\frac{10}{11}}, \frac{10x}{11} + z_{1-\alpha/2} \sqrt{\frac{10}{11}} \right]$$

es un intervalo estimador de nivel $1 - \alpha$. □

1.5. Sobre la distribución a priori uniforme.

Cuando el parámetro θ tiene distribución a priori $\mathcal{U}[a, b]$, esto es $\pi_\theta(t) = \frac{1}{b-a} \mathbf{1}\{t \in [a, b]\}$ el enfoque Bayesiano se simplifica abruptamente.

La fórmula de Bayes para la distribución a posteriori (4) adopta la forma

$$\pi(t|\mathbf{x}) = \frac{L(t|\mathbf{x}) \frac{1}{b-a} \mathbf{1}\{t \in [a, b]\}}{\int L(t|\mathbf{x}) \frac{1}{b-a} \mathbf{1}\{t \in [a, b]\} dt} = \frac{L(t|\mathbf{x}) \mathbf{1}\{t \in [a, b]\}}{\int_a^b L(t|\mathbf{x}) dt}. \quad (11)$$

En palabras, si la distribución a priori del parámetro es uniforme, la densidad de su distribución a posteriori es proporcional a la función de verosimilitud: $\pi(t|\mathbf{x}) \propto L(t|\mathbf{x})$.

Nota Bene. En cierto sentido, que puede precisarse, la distribución $\mathcal{U}[a, b]$ es la *menos informativa* entre todas las distribuciones continuas a valores en $[a, b]$.

En *teoría de la información* la indeterminación de una variable aleatoria X se mide con la *entropía* definida por $H(X) := \mathbb{E}[-\log f(X)]$, donde $f(x)$ es la densidad de probabilidades de la variable aleatoria X . En otros términos

$$H(X) := - \int f(x) \log f(x) dx. \quad (12)$$

Teorema 1.8. Entre todas las variables aleatorias continuas a valores en $[a, b]$ la que maximiza la entropía es la $\mathcal{U}[a, b]$.

Demostración. No se pierde generalidad si se supone que $[a, b] = [0, 1]$. Si $X \sim \mathcal{U}[0, 1]$, entonces

$$H(X) = - \int_0^1 1 \log(1) dx = 0.$$

El resultado se obtiene mostrando que si X es una variable aleatoria continua a valores en el $[0, 1]$, entonces $H(X) \leq 0$.

Es fácil ver que para todo $x > 0$ vale la desigualdad

$$\log(x) \leq x - 1 \quad (13)$$

Poniendo $x = \frac{1}{u}$, $u > 0$, en la desigualdad (13) se obtiene

$$-\log u = \log\left(\frac{1}{u}\right) \leq \frac{1}{u} - 1 \quad (14)$$

La desigualdad (14) se usa para obtener

$$H(X) = - \int_0^1 f(x) \log f(x) dx \leq \int_0^1 f(x) \left(\frac{1}{f(x)} - 1 \right) dx = \int_0^1 1 dx - \int_0^1 f(x) dx = 0.$$

□

Comentario Bibliográfico. Una exposición elemental de la noción de entropía y de las distribuciones menos informativas puede leerse en Pugachev, V.S., (1973). *Introducción a la Teoría de Probabilidades*, Mir, Moscú.

Enfoque Bayesiano generalizado. Si la función de verosimilitud $L(t|\mathbf{x})$ es integrable, i.e., $0 < \int_{-\infty}^{\infty} L(t|\mathbf{x}) dt < \infty$, la expresión

$$\pi(t|\mathbf{x}) := \frac{L(t|\mathbf{x})}{\int_{-\infty}^{\infty} L(t|\mathbf{x}) dt} \quad (15)$$

define una densidad de probabilidades en \mathbb{R} . Por abuso del lenguaje, algunos autores suelen llamarla la densidad a posteriori correspondiente a la distribución a priori “uniforme sobre la recta”¹. No hay ningún problema en utilizar este enfoque siempre que no se pierda de vista que no existe ninguna distribución uniforme sobre regiones de longitud infinita. El enfoque que postula una densidad a posteriori de la forma (15) será llamado *Bayesiano generalizado*.

2. Ejemplos

2.1. Las distribuciones β y el problema del “control de calidad”

Control de calidad. La calidad de un proceso de producción puede medirse por el porcentaje, $100\theta\%$, de artículos defectuosos producidos. Cada artículo producido tiene asociada

¹**Nota histórica:** la denominación para esta a priori impropia se debe a Laplace.

una variable aleatoria de Bernoulli, $X \sim \text{Bernoulli}(\theta)$, cuyo parámetro θ denota la probabilidad de que el artículo sea defectuoso.

El punto de partida del enfoque Bayesiano es la distribución a priori del parámetro. Supongamos que, a priori, $\theta \sim \mathcal{U}(0, 1)$. Se observa una muestra aleatoria $\mathbf{X} = (X_1, \dots, X_n)$ y usando la fórmula de Bayes (4) se obtiene la densidad, $\pi(t|\mathbf{x})$, de la distribución a posteriori de θ dado que $\mathbf{X} = \mathbf{x}$. Cuando la densidad a priori es uniforme la densidad a posteriori es proporcional a la verosimilitud. Por lo tanto,

$$\pi(t|\mathbf{x}) \propto L(t|\mathbf{x}) = t^{k(\mathbf{x})}(1-t)^{n-k(\mathbf{x})}\mathbf{1}\{t \in (0, 1)\}, \quad (16)$$

donde $k(\mathbf{x}) = \sum_{i=1}^n x_i$. De la identidad (16) se concluye que $\theta|\mathbf{X} = \mathbf{x}$ tiene una distribución beta de parámetros $k(\mathbf{x}) + 1$ y $n - k(\mathbf{x}) + 1$. En consecuencia la constante de proporcionalidad será

$$\frac{\Gamma(n+2)}{\Gamma(k(\mathbf{x})+1)\Gamma(n-k(\mathbf{x})+1)} = \frac{(n+1)!}{k(\mathbf{x})!(n-k(\mathbf{x}))!} = (n+1)\binom{n}{k(\mathbf{x})}. \quad (17)$$

Conclusión. Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de volumen n correspondiente a una variable aleatoria $X \sim \text{Bernoulli}(\theta)$. Si la distribución a priori del parámetro θ es uniforme sobre el intervalo $(0, 1)$ y se observa que $\mathbf{X} = \mathbf{x}$, entonces la distribución a posteriori (del parámetro θ) es una $\beta(k+1, n-k+1)$, donde k es la cantidad de éxitos observados. En otras palabras, la densidad de $\theta|\mathbf{X} = \mathbf{x}$ es

$$\pi(t|\mathbf{x}) = (n+1)\binom{n}{k}t^k(1-t)^{n-k}\mathbf{1}\{t \in (0, 1)\}, \quad (18)$$

donde $k = \sum_{i=1}^n x_i$. □

Función de probabilidad marginal. Cuál es la probabilidad de que en una muestra de volumen n se observen exactamente k artículos defectuosos. La cantidad de artículos defectuosos será $N = \sum_{i=1}^n X_i$. Dado que $\theta = t$, las variables X_1, \dots, X_n serán independientes, cada una con distribución de Bernoulli(t) y en tal caso $N \sim \text{Binomial}(n, t)$

$$\mathbb{P}(N = k|t) = \binom{n}{k}t^k(1-t)^{n-k}, \quad k = 0, 1, \dots, n \quad (19)$$

Por lo tanto, condicionando sobre $\theta = t$ y usando la fórmula de probabilidad total, obtenemos que

$$\begin{aligned} \mathbb{P}(N = k) &= \int_0^1 \mathbb{P}(N = k|t)\pi_\theta(t)dt = \int_0^1 \binom{n}{k}t^k(1-t)^{n-k}dt \\ &= \binom{n}{k} \int_0^1 t^k(1-t)^{n-k}dt = \binom{n}{k} \frac{k!(n-k)!}{(n+1)!} \\ &= \frac{1}{n+1} \quad k = 0, 1, \dots, n \end{aligned} \quad (20)$$

En otras palabras, los $n+1$ valores posibles de N son igualmente probables.

Función de probabilidad predictiva Supongamos ahora que en una muestra de volumen n se observaron exactamente k artículos defectuosos. Cuál es la probabilidad p de que un nuevo artículo resulte defectuoso?

Para calcular p usamos la función de probabilidad predictiva obtenida en (5):

$$p = f(1|\mathbf{x}) = \int_0^1 f(1|t)\pi(t|\mathbf{x})dt = \int_0^1 t\pi(t|\mathbf{x})dx = \mathbb{E}[\theta|\mathbf{X} = \mathbf{x}] = \frac{k+1}{n+2}. \quad (21)$$

Esto es, si los primeros n artículos resultaron en k defectuosos, entonces el próximo artículo será defectuoso con probabilidad $(k+1)/(n+2)$.

De la ecuación (21) resulta una descripción alternativa del proceso de producción examinado: Hay una urna que inicialmente contiene una bola blanca y una bola negra. En cada paso se extrae al azar una bola de la urna y se la repone junto con otra del mismo color. Despues de cada extracción la cantidad de bolas del color extraído aumenta una unidad y la cantidad de bolas del color opuesto se mantiene constante. Si de las primeras n bolas elegidas, k fueron blancas, entonces en la urna al momento de la $n+1$ -ésima extracción hay $k+1$ blancas y $n-k+1$ negras, y por lo tanto la siguiente bola será blanca con probabilidad $(k+1)/(n+2)$. Identificando la extracción de una bola blanca con un artículo defectuoso, tenemos una descripción alternativa del modelo original. Esté último se llama *modelo de urna de Polya*.

Estimadores Bayesianos

- Utilizando la esperanza condicional de $\theta|\mathbf{X} = \mathbf{x}$ obtenemos la siguiente estimación

$$\hat{\theta}(\mathbf{x}) = \mathbb{E}[\theta|\mathbf{X} = \mathbf{x}] = \frac{1}{n+2} \left(1 + \sum_{i=1}^n x_i \right). \quad (22)$$

- El estimador máximo a posteriori se obtiene observando que

$$\begin{aligned} \hat{\theta}_{map}(\mathbf{x}) &= \arg \max_{t \in (0,1)} (n+1) \binom{n}{k} t^k (1-t)^{n-k} = \arg \max_{t \in (0,1)} t^k (1-t)^{n-k} \\ &= \arg \max_{t \in (0,1)} \log t^k (1-t)^{n-k} = \arg \max_{t \in (0,1)} (k \log t + (n-k) \log(1-t)) \\ &= \frac{k}{n}, \end{aligned}$$

donde $k = \sum_{i=1}^n x_i$. Por lo tanto,

$$\hat{\theta}_{map}(\mathbf{x}) = \bar{x}. \quad (23)$$

Nota Bene. Notar que

$$\hat{\theta}(\mathbf{x}) = \frac{n}{n+2} \bar{x} + \frac{1}{n+2} = \frac{n}{n+2} \bar{x} + \frac{2}{n+2} \mathbb{E}[\mathcal{U}(0,1)],$$

donde $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Estimación por intervalo Se quiere construir un intervalo estimador (de nivel $1 - \alpha$) para θ sabiendo que en una muestra de volumen n se observaron k artículos defectuosos.

En este caso la ecuación (9) adopta la forma

$$1 - \alpha = \int_a^b \frac{(n+1)!}{k!(n-k)!} t^k (1-t)^{n-k} dt. \quad (24)$$

El problema equivale a encontrar las raíces de un polinomio de grado $n+1$ en las variables a y b y no hay métodos generales para encontrarlas. El problema se puede resolver mediante alguna técnica de cálculo numérico para aproximar raíces de polinomios implementada en un computador. Para $3 \leq n+1 \leq 4$ pueden utilizarse las fórmulas de Tartaglia para resolver ecuaciones de tercer y cuarto grado. Estas fórmulas pueden consultarse en el Tomo 1 del *Análisis matemático* de Rey Pastor.

Cuando $k = 0$ o $k = n$ la ecuación (24) se puede resolver “a mano”: si $k = 0$ la ecuación (24) adopta la forma

$$\begin{aligned} 1 - \alpha &= \int_a^b (n+1)(1-t)^n dt = (n+1) \left(-\frac{(1-t)^{n+1}}{n+1} \Big|_a^b \right) \\ &= (n+1) \left(\frac{(1-a)^{n+1}}{n+1} - \frac{(1-b)^{n+1}}{n+1} \right) \\ &= (1-a)^{n+1} - (1-b)^{n+1}. \end{aligned}$$

Fijado un valor “razonable” de a se puede despejar el valor de b

$$b = 1 - \sqrt[n+1]{(1-a)^{n+1} - (1-\alpha)}, \quad 0 \leq a \leq 1 - \sqrt[n+1]{1-\alpha} \quad (25)$$

Hemos visto que, para $k = 0$ el máximo a posteriori es 0, poniendo $a = 0$ se obtiene $b = 1 - \sqrt[n+1]{\alpha}$. Por lo tanto, el intervalo

$$[0, 1 - \sqrt[n+1]{\alpha}]$$

es un intervalo estimador de nivel $1 - \alpha$.

Ejemplo 2.1. Sea X una variable aleatoria Bernoulli de parámetro θ . A priori se supone que la distribución de θ es uniforme sobre el intervalo $[0, 1]$. Supongamos que una muestra aleatoria de volumen $n = 20$ arroja los siguientes resultados:

$$\mathbf{x} = (0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1)$$

Distribución a posteriori. Como la cantidad de éxitos observados es $k = 11$, tenemos que $\theta | \mathbf{X} = \mathbf{x} \sim \beta(12, 10)$. En otras palabras, la densidad a posteriori es de la forma

$$\pi(t | \mathbf{x}) = \frac{21!}{11!9!} t^{11} (1-t)^9 \mathbf{1}\{t \in [0, 1]\}. \quad (26)$$

En la Figura 1 se muestran los gráficos de la distribución a priori de θ y de la distribución a posteriori de θ vista la muestra.

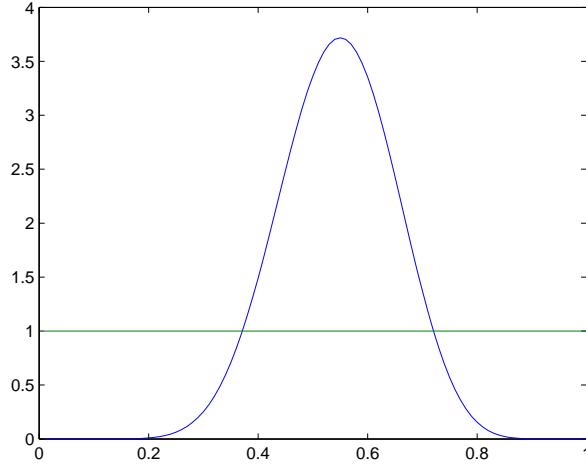


Figura 1: Gráficos de las densidades a priori y a posteriori: en verde el gráfico de la densidad de la distribución $\mathcal{U}[0, 1]$ y en azul el de la distribución $\beta(12, 10)$.

Predicción. ¿Cuál es la probabilidad de que en una nueva muestra de volumen 5 resulten exactamente 2 éxitos?

En primer lugar hay que observar que dado que $\theta = t$ la cantidad de éxitos N en una muestra de volumen 5 tiene distribución Binomial(5, t). Por lo tanto,

$$\mathbb{P}(N = 2|t) = \binom{5}{2} t^2 (1-t)^3 = 10t^2(1-t)^3.$$

Como la densidad a posteriori de θ resultó ser

$$\pi(t|\mathbf{x}) = \frac{21!}{11!9!} t^{11} (1-t)^9 \mathbf{1}\{t \in [0, 1]\},$$

de la fórmula de probabilidad total se deduce que

$$\begin{aligned} \mathbb{P}(N = 2|\mathbf{x}) &= \int_0^1 \mathbb{P}(N = 2|t) f(t|\mathbf{x}) dt = \int_0^1 10t^2(1-t)^3 \frac{21!}{11!9!} t^{11} (1-t)^9 dt \\ &= 10 \frac{21!}{11!9!} \int_0^1 t^{13} (1-t)^{12} dt = 10 \frac{21!}{11!9!} \frac{13!12!}{26!} = \frac{6}{23} = 0.26\dots \end{aligned}$$

Estimadores Bayesianos

1. Esperanza condicional:

$$\hat{\theta} = \mathbb{E}[\theta|\mathbf{X} = \mathbf{x}] = \frac{12}{22} = \frac{6}{11} = 0.5454\dots$$

2. Máximo a posteriori:

$$\hat{\theta}_{map} = \bar{x} = \frac{11}{20} = 0.55.$$

Estimación por intervalo Para construir un intervalo $[a, b]$, de nivel 0.95, para θ podemos resolver las siguientes ecuaciones

$$\int_0^a \frac{21!}{11!9!} t^{11} (1-t)^9 dt = 0.025, \quad \int_0^b \frac{21!}{11!9!} t^{11} (1-t)^9 dt = 0.975.$$

Utilizando una herramienta de cálculo obtenemos que $a = 0.3402$ y $b = 0.7429$. \square

2.2. Normales de varianza conocida y media normal

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una familia normal $\mathcal{N}(\theta, \sigma^2)$, con σ^2 conocido. Supongamos que la distribución a priori del parámetro θ es una normal $\mathcal{N}(\mu, \rho^2)$.

Distribución a posteriori. Por definición, ver (4), la densidad a posteriori de θ , dado que $\mathbf{X} = \mathbf{x}$, queda caracterizada por la relación de proporcionalidad $\pi(t|\mathbf{x}) \propto L(t|\mathbf{x})\pi_\theta(t)$, donde $L(t|\mathbf{x})$ es la función de verosimilitud y $\pi_\theta(t)$ la densidad a priori de θ .

Primero calculamos la función de verosimilitud. De las igualdades

$$\begin{aligned} L(\mu, \sigma^2 | \mathbf{x}) &= \prod_{i=1}^n f(x_i | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right), \end{aligned} \quad (27)$$

donde $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$,² se deduce que

$$L(t|\mathbf{x}) \propto \exp\left(-\frac{n(\bar{x} - t)^2}{2\sigma^2}\right). \quad (28)$$

Por hipótesis, $\theta \sim \mathcal{N}(\mu, \rho^2)$. En consecuencia,

$$\pi_\theta(t) \propto \exp\left(-\frac{(t - \mu)^2}{2\rho^2}\right) \quad (29)$$

De (28) y (29), la densidad a posteriori satisface

$$\pi(t|\mathbf{x}) \propto \exp\left(-\left[\frac{n(\bar{x} - t)^2}{2\sigma^2} + \frac{(t - \mu)^2}{2\rho^2}\right]\right). \quad (30)$$

Completando cuadrados respecto de t se obtiene

$$\frac{n(\bar{x} - t)^2}{2\sigma^2} + \frac{(t - \mu)^2}{2\rho^2} = \frac{n\rho^2 + \sigma^2}{2\sigma^2\rho^2} \left(t - \frac{n\rho^2\bar{x} + \sigma^2\mu}{n\rho^2 + \sigma^2}\right)^2 + \text{otras cosas} \quad (31)$$

²La última igualdad de (27) se obtiene observando que

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2.$$

donde “otras cosas” son expresiones que no dependen de t . En consecuencia,

$$\pi(t|\mathbf{x}) \propto \exp\left(-\frac{n\rho^2 + \sigma^2}{2\sigma^2\rho^2} \left(t - \frac{n\rho^2\bar{x} + \sigma^2\mu}{n\rho^2 + \sigma^2}\right)^2\right). \quad (32)$$

Por lo tanto, la distribución a posteriori de θ dado que $\mathbf{X} = \mathbf{x}$ es una normal

$$\mathcal{N}\left(\frac{n\rho^2\bar{x} + \sigma^2\mu}{n\rho^2 + \sigma^2}, \frac{\sigma^2\rho^2}{n\rho^2 + \sigma^2}\right). \quad (33)$$

Función densidad predictiva. Comenzamos calculando el producto de la densidad condicional de X dado que $\theta = t$ por la densidad a posteriori de θ dado que $\mathbf{X} = \mathbf{x}$:

$$\begin{aligned} f(x|t)\pi(t|\mathbf{x}) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-t)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\rho_*} \exp\left(-\frac{(t-\mu_*)^2}{2\rho_*^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}\rho_*\sigma} \exp\left(-\left[\frac{(x-t)^2}{2\sigma^2} + \frac{(t-\mu_*)^2}{2\rho_*^2}\right]\right), \end{aligned} \quad (34)$$

donde μ_* y ρ_*^2 son la media y la varianza de la distribución a posteriori de θ dado que $\mathbf{X} = \mathbf{x}$

$$\mu_* = \frac{n\rho^2\bar{x} + \sigma^2\mu}{n\rho^2 + \sigma^2} \quad \text{y} \quad \rho_*^2 = \frac{\sigma^2\rho^2}{n\rho^2 + \sigma^2} \quad (35)$$

Con un poco de paciencia, puede verse que

$$\frac{(x-t)^2}{2\sigma^2} + \frac{(t-\mu_*)^2}{2\rho_*^2} = \frac{\rho_*^2 + \sigma^2}{2\sigma^2\rho_*^2} \left(t - \frac{\rho_*^2x + \sigma^2\mu_*}{\rho_*^2 + \sigma^2}\right)^2 + \frac{(x-\mu_*)^2}{2(\rho_*^2 + \sigma^2)} \quad (36)$$

En consecuencia,

$$\begin{aligned} f(x|t)\pi(t|\mathbf{x}) &= \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{2\pi}\rho_*} \exp\left(-\left[\frac{\rho_*^2 + \sigma^2}{2\sigma^2\rho_*^2} \left(t - \frac{\rho_*^2x + \sigma^2\mu_*}{\rho_*^2 + \sigma^2}\right)^2 + \frac{(x-\mu_*)^2}{2(\rho_*^2 + \sigma^2)}\right]\right) \\ &= \frac{1}{\sqrt{2\pi(\rho_*^2 + \sigma^2)}} \exp\left(-\frac{(x-\mu_*)^2}{2(\rho_*^2 + \sigma^2)}\right) \\ &\times \frac{1}{\sqrt{2\pi\frac{\rho_*^2\sigma^2}{\rho_*^2 + \sigma^2}}} \exp\left(-\frac{\rho_*^2 + \sigma^2}{2\sigma^2\rho_*^2} \left(t - \frac{\rho_*^2x + \sigma^2\mu_*}{\rho_*^2 + \sigma^2}\right)^2\right). \end{aligned} \quad (37)$$

Integrando respecto de t , ambos lados de identidad (37), obtenemos la expresión de la densidad predictiva

$$f(x|\mathbf{x}) = \int f(x|t)\pi(t|\mathbf{x})dt = \frac{1}{\sqrt{2\pi(\rho_*^2 + \sigma^2)}} \exp\left(-\frac{(x-\mu_*)^2}{2(\rho_*^2 + \sigma^2)}\right). \quad (38)$$

En otras palabras, la distribución de la variable aleatoria X dado que $\mathbf{X} = \mathbf{x}$, es una normal de media μ_* y varianza $\sigma^2 + \rho_*^2$. El resultado obtenido nos permite calcular todas las probabilidades de la forma $\mathbb{P}(X \in A|\mathbf{X} = \mathbf{x})$.

Estimadores Bayesianos. En este caso, como el máximo de la normal se alcanza en la media ambos estimadores coinciden:

$$\hat{\theta} = \frac{n\rho^2\bar{x} + \sigma^2\mu}{n\rho^2 + \sigma^2}. \quad (39)$$

Nota Bene. Note que

$$\hat{\theta} = \frac{n\rho^2}{n\rho^2 + \sigma^2}\bar{x} + \frac{\sigma^2}{n\rho^2 + \sigma^2}\mu = \frac{n\rho^2}{n\rho^2 + \sigma^2}\bar{x} + \frac{\sigma^2}{n\rho^2 + \sigma^2}\mathbb{E}[\mathcal{N}(\mu, \rho^2)] \quad (40)$$

Estimación por intervalo. En lo que sigue construiremos un intervalo estimador de nivel $1 - \alpha$ para θ sabiendo que $\mathbf{X} = \mathbf{x}$. Sabemos que $\theta|\mathbf{X} = \mathbf{x}$ se distribuye como una normal de media μ_* y varianza ρ_*^2 . Proponiendo un intervalo centrado en la media μ_* de la forma

$$[\mu_* - \epsilon, \mu_* + \epsilon] \quad (41)$$

y usando la simetría de la normal con respecto a su media, el problema se reduce a encontrar el valor de ϵ que resuelve la ecuación siguiente

$$1 - \frac{\alpha}{2} = \mathbb{P}(\theta \leq \mu_* + \epsilon | \mathbf{X} = \mathbf{x}) = \mathbb{P}\left(\frac{\theta - \mu_*}{\rho_*} \leq \frac{\epsilon}{\rho_*} \mid \mathbf{X} = \mathbf{x}\right) = \Phi\left(\frac{\epsilon}{\rho_*}\right). \quad (42)$$

En consecuencia,

$$\epsilon = \rho_*\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \sqrt{\frac{\sigma^2\rho^2}{n\rho^2 + \sigma^2}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \frac{\sigma\rho}{\sqrt{n\rho^2 + \sigma^2}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (43)$$

Por lo tanto, el intervalo

$$\left[\frac{n\rho^2\bar{x} + \sigma^2\mu}{n\rho^2 + \sigma^2} - \frac{\sigma\rho}{\sqrt{n\rho^2 + \sigma^2}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \frac{n\rho^2\bar{x} + \sigma^2\mu}{n\rho^2 + \sigma^2} + \frac{\sigma\rho}{\sqrt{n\rho^2 + \sigma^2}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right] \quad (44)$$

es un intervalo estimador de nivel $1 - \alpha$ para θ sabiendo que $\mathbf{X} = \mathbf{x}$. Note que la longitud del intervalo no depende los valores arrojados por la muestra y es del orden de $\frac{1}{\sqrt{n}}$.

Curva peligrosa. Para una muestra de una $\mathcal{N}(\theta, \sigma^2)$ con distribución a priori para θ de la forma $\mathcal{N}(\mu, \rho^2)$ obtuvimos que la distribución a posteriori satisface

$$f(t|\mathbf{x}) \propto \exp\left(-\frac{n\rho^2 + \sigma^2}{2\sigma^2\rho^2}\left(t - \frac{n\rho^2\bar{x} + \sigma^2\mu}{n\rho^2 + \sigma^2}\right)^2\right). \quad (45)$$

A medida que aumentamos el valor de ρ^2 la información contenida en la distribución a priori se va “destruyendo” y la densidad a posteriori se va aproximando a la densidad de una normal de media \bar{x} y varianza σ^2/n :

$$\lim_{\rho^2 \rightarrow \infty} f(t|\mathbf{x}) \propto \exp\left(-\frac{n(t - \bar{x})^2}{2\sigma^2}\right) \propto L_t(\mathbf{x}). \quad (46)$$

En palabras informales y poco rigurosas, si se destruye la información contenida en la distribución a priori $\mathcal{N}(\mu, \rho^2)$ mediante el procedimiento de hacer $\rho^2 \rightarrow \infty$ se obtiene una densidad de probabilidades proporcional a la verosimilitud. Vale decir, en el caso límite se obtiene el *enfoque Bayesiano generalizado*. Desde esta perspectiva, el enfoque Bayesiano generalizado puede interpretarse como una metodología orientada a destruir toda la información contenida en las distribuciones a priori del parámetro.

Ejemplo 2.2. Se tiene la siguiente muestra aleatoria de volumen $n = 10$ de una población $\mathcal{N}(\theta, 1)$

$$\begin{array}{ccccc} 2.0135 & 0.9233 & 0.0935 & 0.0907 & 0.3909 \\ 0.3781 & -1.9313 & -0.8401 & 3.4864 & -0.6258 \end{array}$$

Si, a priori, suponemos que $\theta \sim \mathcal{N}(0, 1)$, entonces la distribución a posteriori de θ es una normal, ver (33), $\mathcal{N}\left(\frac{10\bar{x}}{11}, \frac{1}{11}\right)$. Observando la muestra se obtiene que $\bar{x} = 0.3979$. Por lo tanto, la distribución a posteriori del parámetro es una normal $\mathcal{N}\left(\frac{3.979}{11}, \frac{1}{11}\right)$.

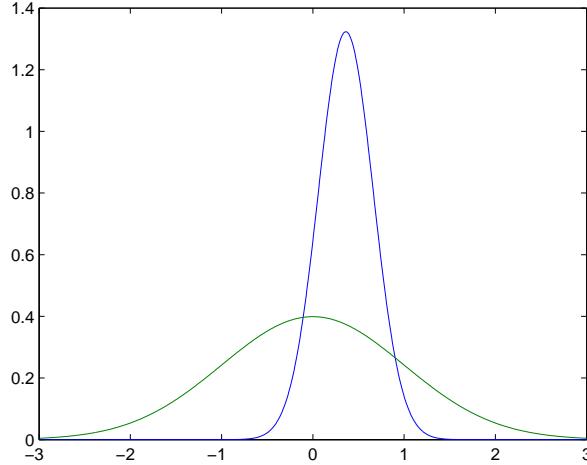


Figura 2: Gráficos de las densidades a priori (en verde) y a posteriori (en azul).

Como la moda y la media de la distribución normal coinciden, el estimador puntual Bayesiano resulta ser $\hat{\theta} = 3.979/11 = 0.3617\dots$

Utilizando la tabla de la normal estándar puede verse que $I = [-0.2292, 0.9527]$ es un intervalo de nivel 0.95.

Etcétera... □

2.3. Distribuciones Poisson con a priori Gamma

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una distribución Poisson de parámetro θ , $\theta > 0$. Supongamos que la distribución a priori del parámetro θ es una Gamma de parámetros ν y λ . Esto es, la densidad a priori del parámetro es de la forma

$$\pi_\theta(t) \propto t^{\nu-1} e^{-\lambda t} \mathbf{1}\{t > 0\} \quad (47)$$

Distribución a posteriori. La densidad a posteriori de θ , dado que $\mathbf{X} = \mathbf{x}$, queda caracterizada por la relación de proporcionalidad $\pi(t|\mathbf{x}) \propto L(t|\mathbf{x})\pi_\theta(t)$, donde $L(t|\mathbf{x})$ es la función de verosimilitud y $\pi_\theta(t)$ es la densidad a priori de θ . En este caso la función de verosimilitud es de la forma

$$L(t|\mathbf{x}) \propto e^{-nt} t^{\sum_{i=1}^n x_i}. \quad (48)$$

De (47) y (48) se deduce que la densidad a posteriori de θ dado que $\mathbf{X} = \mathbf{x}$ satisface

$$\pi(t|\mathbf{x}) \propto e^{-nt} t^{\sum_{i=1}^n x_i} t^{\nu-1} e^{-\lambda t} \mathbf{1}\{t > 0\} = t^{\sum_{i=1}^n x_i + \nu - 1} e^{-(n+\lambda)t} \mathbf{1}\{t > 0\}. \quad (49)$$

Por lo tanto, la distribución a posteriori de θ dado que $\mathbf{X} = \mathbf{x}$ es una Gamma

$$\Gamma\left(\sum_{i=1}^n x_i + \nu, n + \lambda\right).$$

Estimadores Bayesianos.

- Utilizando la esperanza condicional de $\theta|\mathbf{X} = \mathbf{x}$ obtenemos la siguiente estimación.

$$\hat{\theta} = \mathbb{E}[\theta|\mathbf{X} = \mathbf{x}] = \frac{\sum_{i=1}^n x_i + \nu}{n + \lambda} \quad (50)$$

- La estimación por máximo a posteriori se obtiene observando que

$$\arg \max_{t>0} t^a e^{-bt} = \arg \max_{t>0} \log t^a e^{-bt} = \arg \max_{t>0} (a \log t - bt) = \frac{b}{a}.$$

Por lo tanto,

$$\hat{\theta}_{map} = \frac{\sum_{i=1}^n x_i + \nu - 1}{n + \lambda}. \quad (51)$$

Nota Bene. Notar que

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=1}^n x_i + \nu}{n + \lambda} = \frac{n}{n + \lambda} \left(\frac{\sum_{i=1}^n x_i}{n} \right) + \frac{\lambda}{n + \lambda} \left(\frac{\nu}{\lambda} \right) \\ &= \frac{n}{n + \lambda} \bar{x} + \frac{\lambda}{n + \lambda} \mathbb{E}[\Gamma(\nu, \lambda)]. \end{aligned} \quad (52)$$

Función de probabilidad predictiva. El producto de la probabilidad condicional de X dado que $\theta = t$ por la densidad a posteriori de θ dado que $\mathbf{X} = \mathbf{x}$:

$$\begin{aligned} f(x|t)\pi(t|\mathbf{x}) &= e^{-t} \frac{t^x}{x!} \frac{(n + \lambda)^{\nu(\mathbf{x})}}{\Gamma(\nu(\mathbf{x}))} t^{\nu(\mathbf{x})-1} e^{-(n+\lambda)t} \mathbf{1}\{t > 0\} \\ &= \frac{(n + \lambda)^{\nu(\mathbf{x})}}{x! \Gamma(\nu(\mathbf{x}))} t^{\nu(\mathbf{x})+x-1} e^{-(n+\lambda+1)t} \mathbf{1}\{t > 0\}, \end{aligned} \quad (53)$$

donde $\nu(\mathbf{x}) = \sum_{i=1}^n x_i + \nu$. Integrando respecto de t ambos lados de la identidad (53), obtenemos la expresión de la función de probabilidad incondicional (o predictiva)

$$\begin{aligned}
f(x|\mathbf{x}) &= \frac{(n+\lambda)^{\nu(\mathbf{x})}}{x!\Gamma(\nu(\mathbf{x}))} \int_0^\infty t^{\nu(\mathbf{x})+x-1} e^{-(n+\lambda+1)t} dt \\
&= \frac{(n+\lambda)^{\nu(\mathbf{x})}}{x!\Gamma(\nu(\mathbf{x}))} \frac{\Gamma(\nu(\mathbf{x})+x)}{(n+\lambda+1)^{\nu(\mathbf{x})+x}} \\
&= \frac{\Gamma(\nu(\mathbf{x})+x)}{\Gamma(\nu(\mathbf{x}))x!} \frac{(n+\lambda)^{\nu(\mathbf{x})}}{(n+\lambda+1)^{\nu(\mathbf{x})+x}} \\
&= \frac{\Gamma(\nu(\mathbf{x})+x)}{\Gamma(\nu(\mathbf{x}))x!} \left(\frac{1}{n+\lambda+1} \right)^x \left(\frac{n+\lambda}{n+\lambda+1} \right)^{\nu(\mathbf{x})}.
\end{aligned} \tag{54}$$

Una expresión que con un poco de paciencia (o una computadora a la mano) se puede calcular para cada valor de x .

Caso $\nu \in \mathbb{N}$. En este caso la expresión para la función de probabilidad incondicional (54) adopta la forma

$$\begin{aligned}
f(x|\mathbf{x}) &= \frac{(\nu(\mathbf{x})+x-1)!}{(\nu(\mathbf{x})-1)!x!} \left(\frac{1}{n+\lambda+1} \right)^x \left(\frac{n+\lambda}{n+\lambda+1} \right)^{\nu(\mathbf{x})} \\
&= \binom{\nu(\mathbf{x})+x-1}{\nu(\mathbf{x})-1} \left(\frac{1}{n+\lambda+1} \right)^x \left(\frac{n+\lambda}{n+\lambda+1} \right)^{\nu(\mathbf{x})}.
\end{aligned} \tag{55}$$

La expresión (55) para la función de probabilidad condicional $f(x|\mathbf{x})$ admite la siguiente interpretación probabilística: *Dado que $\mathbf{X} = \mathbf{x}$, la probabilidad incondicional de que la variable Poisson asuma el valor x es igual a la probabilidad de que en una sucesión de ensayos Bernoulli independientes de parámetro $\frac{n+\lambda}{n+\lambda+1}$ el $\nu(\mathbf{x})$ -ésimo éxito ocurra en el $(\nu(\mathbf{x})+x)$ -ésimo ensayo.*

Estimación por intervalo. Dado que $\mathbf{X} = \mathbf{x}$, podemos construir un intervalo estimador de nivel $1 - \alpha$ para θ observando que

$$2(n+\lambda)\theta \sim \Gamma\left(\frac{2\nu(\mathbf{x})}{2}, \frac{1}{2}\right).$$

Si además $\nu \in \mathbb{N}$, entonces

$$2(n+\lambda)\theta \sim \chi^2_{2\nu(\mathbf{x})}.$$

En tal caso,

$$\mathbb{P}\left(2(n+\lambda)\theta \in \left[\chi^2_{2\nu(\mathbf{x}),\alpha/2}, \chi^2_{2\nu(\mathbf{x}),1-\alpha/2}\right]\right) = 1 - \alpha.$$

Por lo tanto, si $\nu \in \mathbb{N}$ y sabiendo que $\mathbf{X} = \mathbf{x}$ el intervalo

$$\left[\frac{\chi^2_{2\nu(\mathbf{x}),\alpha/2}}{2(n+\lambda)}, \frac{\chi^2_{2\nu(\mathbf{x}),1-\alpha/2}}{2(n+\lambda)} \right],$$

donde $\nu(\mathbf{x}) = \sum_{i=1}^n x_i + \nu$, es un intervalo estimador de nivel $1 - \alpha$ para θ .

Ejemplo 2.3. La cantidad de errores de tipeo por hoja que comete una secretaria profesional puede modelarse con una distribución de Poisson de parámetro θ (¿Por qué?). A priori, se supone que el parámetro θ sigue una distribución exponencial de intensidad 1 (Esta hipótesis sobre la distribución de θ es la menos informativa si se supone que la media de la distribución es 1). Se analizan 10 hojas tipeadas por la mencionada secretaria y resulta que la cantidad de errores por página es

1 3 3 3 4 6 3 2 2 2

Si la secretaria tipea una nueva hoja, cuál es la probabilidad de que cometa como máximo un error?

Solución. Para resolver este problema utilizaremos la función de probabilidad predictiva. De acuerdo con (54), como la distribución a priori de θ es una $\text{Exp}(1) = \Gamma(1, 1)$, dicha función es de la forma

$$f(x|\mathbf{x}) = \binom{\nu(\mathbf{x}) + x - 1}{\nu(\mathbf{x}) - 1} \left(\frac{1}{n + \lambda + 1} \right)^x \left(\frac{n + \lambda}{n + \lambda + 1} \right)^{\nu(\mathbf{x})} = \binom{29 + x}{29} \left(\frac{1}{12} \right)^x \left(\frac{11}{12} \right)^{30},$$

debido a que $n = 10$, $\nu(\mathbf{x}) = \sum_{i=1}^n x_i + 1 = 30$ y $\lambda = 1$. Por lo tanto, la probabilidad de que la secretaria cometa como máximo un error al tipear una nueva hoja será

$$\begin{aligned} f(0|\mathbf{x}) + f(1|\mathbf{x}) &= \binom{29}{29} \left(\frac{1}{12} \right)^0 \left(\frac{11}{12} \right)^{30} + \binom{30}{29} \left(\frac{1}{12} \right)^1 \left(\frac{11}{12} \right)^{30} \\ &= \left(\frac{11}{12} \right)^{30} \left(1 + 30 \left(\frac{1}{12} \right) \right) = \left(\frac{11}{12} \right)^{30} \left(\frac{7}{2} \right) = 0.257\dots \end{aligned}$$

□

3. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Bolfarine, H., Sandoval, M. C.: Introdução à Inferência Estatística. SBM, Rio de Janeiro. (2001)
2. Borovkov, A. A.: Estadística matemática. Mir, Moscú. (1984)
3. Hoel P. G.: Introducción a la estadística matemática. Ariel, Barcelona. (1980)
4. Pugachev, V. S.: Introducción a la Teoría de Probabilidades. Mir, Moscú. (1973)
5. Robert, C. P.: The Bayesian Choice. Springer, New York. (2007)
6. Ross, S. M.: Introduction to Probability and Statistics for Engineers and Scientists. Elsevier Academic Press, San Diego. (2004)